Accepted for publication in the Journal of Experimental Psychology: General

Effects of Individual Differences in Practice on the Reliability and Validity of Three-Minute "Squared" Cognitive Tests

Alexander P. Burgoyne[1,2], Richard Pak[3], Christopher Draheim[4], Ciara Sibley[4], Joseph T. Coyne[4], S. R. Melick[5], & Randall W. Engle[2]

[1]Human Resources Research Organization (HumRRO)

[2]Georgia Institute of Technology

[3]Clemson University

[4]U.S. Naval Research Laboratory, Washington, D.C.

[5]Naval Aerospace Medical Institute, Pensacola, FL

Author Note. Data for Study 1 and R code are openly available at https://osf.io/d5bzy/. Data for Study 2 are kept and protected by the U.S. Navy (specifically, the Naval Aerospace Medical Institute and the Naval Research Laboratory) and can only be shared if the requester can demonstrate to all relevant parties that required data security protocols will be adhered to. The results reported in this manuscript were shared at the Psychonomic Society annual meeting as well as the Society for Industrial and Organizational Psychology annual conference. Task downloads are available at https://englelab.gatech.edu/squaredtasks.html

Alexander P. Burgoyne https://orcid.org/0000-0002-2651-3782

Richard Pak https://orcid.org/0000-0001-9145-6991

Ciara Sibley https://orcid.org/0000-0002-3450-3242

Christopher Draheim https://orcid.org/0000-0002-5224-1389

Joseph T. Coyne https://orcid.org/0000-0001-6461-0148

Randall W. Engle https://orcid.org/0000-0003-0816-7406

Funding. This work was supported by Office of Naval Research grant N00014-22-S-F002 to Randall W. Engle (PI) and Alexander P. Burgoyne (Co-PI).

Abstract

Cognitive ability tests are often administered in settings where test-takers have different amounts of prior experience. We examined whether this heterogeneity undermines test validity using three-minute "Squared" tests of attention control. Across two studies (N > 500 undergraduates; N > 200 U.S. Naval Flight Students), we found that attention control measures remained valid predictors of fluid intelligence, working memory capacity, and performance in Naval aviator training—even when participants had vastly different amounts of practice. In Study 1, participants completed up to 30 practice sessions. We used a novel resampling simulation to model uniform, positively skewed, negatively skewed, and bimodal practice distributions within the sample. Although performance improved with practice, different test-takers having different amounts of practice had minimal impact on criterion-related validity. Furthermore, between-subjects variability increased with practice. Goal reminders designed to reduce cognitive load did not diminish ability-related differences in performance. Study 2 corroborated these findings: attention control predicted Naval flight school grades and provided significant incremental validity beyond the Academic Qualification Rating stanine score, the Navy's selection composite score for predicting academic success, even after practice. We discuss the implications of these results from a theoretical perspective, focusing on theories of skill acquisition and cognitive determinants of performance, as well as from an applied perspective, considering how this approach can inform personnel selection procedures.

Keywords: *Practice effects; goal reminders; attention control; skill acquisition; personnel selection; human performance*

Effects of Individual Differences in Practice on the Reliability and Validity of Three-Minute "Squared" Cognitive Tests

Two individuals take a cognitive test that will determine which one gets a job. One has practiced an online version of the test repeatedly and the other has not practiced the test at all. What are the consequences of different people having different amounts of practice on selection tests?

Any test that is used for personnel selection must contend with *practice effects*: applicants can study or train to try to increase their scores, and in turn, improve their odds of being selected. Although the effects of practice on crystallized cognitive ability tests are well-known (Adesope et al., 2017; Hausknecht et al., 2007), relatively little is known about their effects on tests of fluid cognitive abilities. The purpose of this study was to investigate exactly how practice alters the psychometric properties and validity of attention control tests, particularly under "real-world" conditions in which not every test-taker has the same amount of experience or prior exposure to the test.

Attention Control

The emphasis of this work is on tests of *attention control*, the domain-general ability to maintain focus on task-relevant information over time despite distractions, interruptions, and interference (Burgoyne & Engle, 2020). Attention control is an umbrella term that encompasses executive functions including inhibition, memory updating, and mental set shifting (Miyake & Friedman, 2012). Attention control is closely related to the "central executive" in Baddeley's (1992) model of working memory capacity; it interacts with short-term and long-term memory in service of goal-directed cognition. Individual differences in the ability to control attention predict a range of important outcomes, including working memory capacity, fluid intelligence, sensory

discrimination ability, emotion regulation, academic achievement, and job performance (for a review, see Mashburn et al., 2023). For example, Burgoyne et al. (2024) found that measures of attention control improved the prediction of air traffic controller training performance above and beyond acquired knowledge tests currently used by the U.S. Navy, and Coyne et al. (2024) found a similar pattern of results in a sample of U.S. Navy flight students.

Measuring individual differences in attention control has proven challenging, as many tasks suffer from the psychometric effects of poor signal-to-noise ratios, susceptibility to speed-accuracy tradeoffs, and the use of difference scores (Draheim et al., 2019; Hedge et al., 2018; Rouder et al., 2023). An exhaustive review of these issues is beyond the scope of the present article; we direct the interested reader to Draheim et al. (2019), Hedge et al. (2018), and Rouder et al. (2023).

Recently, however, a trio of three-minute attention control tests appears to have overcome some of these limitations (Burgoyne et al., 2023). The three-minute "Squared" tasks work by adding a twist to classic attention paradigms, such as the Stroop, Flanker, and Simon tasks: they incorporate conflict in the target stimulus *and* in the response options, making them conflict tasks that require inhibition, updating, and set shifting. They are called the Squared tasks because conflict is interactive: while there are main effects of conflict (i.e., incongruency) in the stimulus and response options, there is also a significant interaction between them that influences trial-level performance (Burgoyne et al., 2023; see the Supplemental Materials). Furthermore, they use a timer and points system instead of a response time difference score, improving their reliability.

In an online study and an in-lab study (combined $N > 600$; Burgoyne et al., 2023), the three Squared tests of attention control had a high degree of internal consistency ($\geq .93$),

adequate test-retest reliability (*r*s of .75, .74, and .53), and strong correlations with established measures of attention control (e.g., *r*s of .80 and .81 at the latent level). In turn, an attention control factor identified by the three Squared tasks was highly predictive of performance in synthetic work multitasking paradigms and explained most of the covariation between fluid intelligence, working memory capacity, and processing speed.

Practice Effects and Skill Acquisition

If cognitive ability tests are used for personnel selection, people are likely to practice them. Test preparation materials are widely available for largely knowledge-based tests such as the SAT, GRE, and MCAT (e.g., Lurie, 2010). It is less well-known, however, that even tests of psychomotor and spatial abilities used by the military have been reverse-engineered and made available online, raising concerns about their future reliability and validity (Coyne et al., 2020; Sibley et al., 2023). The ramifications of these practice effects—or *different amounts of practice*, where some test-takers have more experience than others—on the validity of attention control measures is essentially unknown.

Practice effects have implications for theories of skill acquisition and our understanding of 'what the tests test' (Boring, 1961). In their classic theory of skill acquisition, Fitts and Posner (1967) proposed that in the early stages of learning a new skill, performance is effortful and cognitively demanding, but with increasing experience, consistent relations between inputs and outputs can be learned such that performance becomes increasingly automatic. That is, stimulus-response relationships become strengthened such that the presence of a given stimulus seems to automatically elicit the learned response. Similarly, Schneider and Shiffrin (1977) discuss automatic processes as those that can operate without attention and are thus unaffected by capacity limitations. Clearly, this raises a concern for tests of attention control: If performance on

the Squared tests can be routinized through extensive practice, then it may be governed less by attention control processes as a result.

Several studies have found that predictors of performance change as a function of practice. Fleishman and colleagues conducted a program of research on this topic, often using complex motor coordination tasks as their testbed (e.g., stick-and-rudder or rotary pursuit; for a review, see Fleishman [1972]). For example, Fleishman and Hempel (1954) gave subjects 64 trials of practice on a stick-and-rudder coordination task as well as 18 reference tests and then used factor analysis to estimate the variance explained by different cognitive predictors across stages of practice. Some predictors (e.g., spatial relations, visualization) explained less variance as practice progressed, while other increasingly task-specific predictors (e.g., complex coordination, psychomotor coordination, and rate of movement) explained more variance with practice, ultimately accounting for around 75% of the variance at the end of training. As another example, Fleishman and Rich (1963) found that as subjects practiced a two-handed coordination task, spatial orientation was a dominant predictor early on but was ultimately reduced to non-significance, while kinesthetic sensitivity demonstrated the opposite pattern; it did not predict performance early on but rose to a moderate-in-magnitude correlation by the end of the practice trials.

Decades later, Ackerman (1988) offered a theory of ability-performance relations, which argued that performance during the initial stages of learning is primarily governed by general cognitive abilities, but as learning occurs, the role of general cognitive ability is reduced. The prediction that follows is that differences across low- and high-cognitive ability individuals should decrease in magnitude over time, leading to restriction of range (i.e., the "convergence hypothesis," Schmidt et al., 1988, or the "circumvention of limits hypothesis," Hambrick et al.,

2024). In a real-world test of the circumvention of limits hypothesis, Hambrick et al. (2024) found that the validity of cognitive ability for predicting job-specific performance remained stable in a sample of over 10,000 soldiers, even after decades on the job. Evidence for convergence was negligible. It is possible that the changing nature of work over time prevents full routinization of on-the-job task performance, therefore preserving ability-related differences, although this hypothesis warrants empirical substantiation.

Ackerman (1988) further argued that during the intermediate stage of learning, perceptual speed should increase in predictive validity, and during the final stage of learning, the role of perceptual speed should decline while the predictive validity of psychomotor ability should increase. Both restriction of range and changing cognitive mechanisms could alter the validity of a measure. Therefore, assessing the effects of extensive practice is relevant not only to applied issues but also our understanding of the mechanisms underpinning differences in human learning and performance.

Ackerman (1988) conducted several in-lab experiments using tasks that varied in their complexity and stimulus-response consistency to test his predictions, finding mixed support. Throughout the manuscript, Ackerman (1988) claimed the evidence supported many of his theory's predictions, but from a statistical standpoint, this was not the case. For example, Experiment 1 used a choice reaction time task that included a simple training phase and a slightly more complex transfer phase. It was predicted that general cognitive ability would decrease in predictive validity within each phase, but the associated *p*-values were not significant. In the transfer phase, it was claimed that "Perceptual Speed regained influence," (p. 299), but again, the *p*-value was not significant. As another example, in Experiment 2, Ackerman (1988) claimed that the perceptual speed-performance correlation "started low but increased as skills were acquired"

(p. 300), but the *p*-value was reported as $p > .05$. In Experiment 5, Ackerman (1988) claimed that "The ability-performance correlations offer further support for the stated predictions" (p. 304), one of which was that that perceptual speed would show a declining relation to performance later in practice—the associated *p*-value was greater than .05. In Experiment 7, Ackerman (1988) claimed that "there was a trend for Perceptual Speed ability to follow the expected pattern of increasing and then decreasing correlations" (p. 307), yet he also stated that the trend was not statistically significant and did not report the *p*-value. In Experiment 8, Ackerman (1988) stated that the role of reasoning ability "declined with practice, as predicted by the theory" (p. 309), yet the *p*-value exceeded .05. In the same experiment, for perceptual speed and psychomotor ability, Ackerman (1988) claimed that the two cognitive abilities "did indeed follow the expected patterns of results" (p. 309), yet the reported statistical test for perceptual speed was not significant. Collectively, these examples demonstrate a systematic tendency to interpret null results as confirmatory evidence for the presence of an effect, which not only violates conventional standards of statistical inference but also may have led subsequent researchers to accept the theory's empirical support as stronger than it actually was.

More recently, several researchers have examined how practice affects the statistical properties of attention- and memory-related cognitive tests. Xu et al. (2018) examined the stability of visual working memory capacity over repeated testing by having 79 subjects complete a visual-arrays change detection task 30 times (once per day), plus another attempt 30 days later. Subjects improved with practice, by around one standard deviation from the first to the last attempt. Individual differences in performance were relatively stable; the correlation between first and last attempts was $r = .59$ [.42, .71]. Interestingly, the amount of between-subjects variability (*SD*) increased from .85 to 1.08 over the course of practice. Thus, in contrast

to theories that predict range restriction with increasing practice, Xu et al. (2018) observed the opposite result: range *enhancement*. Although Xu et al. did not test this finding for statistical significance, we assessed the correlation between the attempt number and *SD* reported in their Table 2 and found a strong positive correlation, $r = .78$, $p < .001$, indicating that between-subjects variability significantly increased with practice.

Another compelling example of a study of practice effects was reported by Paap et al. (2014). Paap and colleagues had participants complete 20 sessions of practice on the Flanker task from the Attention Network Test, totaling over 4,000 trials per participant, with one participant going on to complete 100 sessions (over 20,000 trials!). Across the first 20 sessions, there was a substantial reduction in the flanker conflict effect, and no clear evidence that a plateau had been reached. This reduction in the flanker effect continued through 100 sessions for one motivated participant, albeit with diminishing returns. This provides compelling evidence that performance on relatively simple, attention-demanding tasks can be characterized by a very long skill acquisition curve, while reinforcing the massive amounts of practice that may be required to reach plateau.

As another example, Whitehead et al. (2019) had participants practice traditional Stroop, Flanker, and Simon tasks for around 1 hour, totaling 960 trials per task. Their analyses focused on whether error-related slowing and the sequential-congruency effect were consistent within individuals across tasks. They found significant correlations across tasks on error-related slowing, but not for the sequential-congruency effect. In another well-powered investigation, Hedge et al. (2018) gave participants 720 trials of practice on the traditional Flanker and Stroop tasks as part of a broader investigation of conflict effects across tasks. In their Supplemental Material D, they report the estimated number of trials per condition required to obtain stable

reliability estimates for response control measures, which generally ranged between 90-150 trials for Flanker and Stroop error costs and RT costs.

In summary, several studies have conducted extensive investigations into practice effects on attention-demanding tasks, but few have included real-world criterion measures, and to our knowledge, not one has investigated what happens when different people have different amounts of practice. These represent critical gaps in the literature that our two studies sought to address.

The Present Studies

In the present studies, we included criterion measures to assess how practice affects the validity of the measures and their relations to other cognitive constructs. In addition, we investigated what happens when different people have different amounts of practice, which is likely to be the case in high-stakes testing scenarios. Indeed, histograms of practice often resemble a positively skewed distribution; most performers have relatively little practice but there is a "long tail" of individuals who have logged progressively more practice than their peers (Hambrick et al., 2014; Gobet & Campitelli, 2007).

In Study 1, we had over 500 participants repeatedly practice attention control tests to examine the effects of varying amounts of practice on the psychometrics of the attention control measures. We also examined how these changes affected the measures' relations with fluid intelligence and working memory capacity. We chose a battery of fluid intelligence and working memory capacity tests as our criterion measures because they reflect the ability to solve novel problems and the ability to keep memoranda in an active state amidst interference, respectively—two sets of cognitive functions that are critical for higher-order cognition. We have argued that attention control largely explains (i.e., mediates) the strong relationship

observed between fluid intelligence and working memory capacity (Engle et al., 1999; Burgoyne & Engle, 2020; Burgoyne et al., 2022; Draheim et al., 2021).

In Study 2, we evaluated the impact of practice in a U.S. military population using real training outcome data. We had a sample of over 200 Student Naval Aviators and Student Naval Flight Officers repeatedly practice an attention control test to assess the impact of practice (and different amounts of practice) on criterion-related validity. Our criterion measures included academic performance and flight performance from the Naval Introductory Flight Evaluation (NIFE) training program. We reasoned that attention control would be important in academic contexts and in dynamic, in-flight contexts. We also tested for incremental validity against the Academic Qualification Rating (AQR), which is one of the measures used to screen and select Naval Flight Students for training. Studies 1 and 2 are complementary in that the first allowed us to investigate more practice attempts in a larger sample, while the second allowed us to extend our approach to an applied context with a highly specialized group of aviators.

Analytical Approach for Varying Amounts of Practice

In Study 1, subjects were randomly assigned to either practice the attention control tests Stroop Squared, Flanker Squared, and Simon Squared 10 times each (i.e., 30 attempts total), or to practice a new attention control task called "Conflict Cubed" 30 times. In total, subjects spent 45 minutes completing test trials on their assigned tasks (i.e., practicing). We estimated the effect of subjects having different amounts of practice by using a novel attempt-sampling approach: we pseudo-randomly sampled scores from different practice attempts for each subject, iterating over this sampling procedure 1,000 times for each simulated distribution of practice. To be clear, this procedure assigns amounts of practice at random, independent of participant ability, motivation,

or coaching. In high-stakes settings, practice is likely to be correlated with these characteristics, a point we return to in the Discussion.

The attempt-sampling approach allowed us to assess the effects of the following distributions of amounts of practice within a sample: 1) uniform distribution (i.e., each participant has an equal likelihood to have any amount of practice); 2) positively-skewed practice (i.e., most participants have little practice, but some have a lot); 3) negatively-skewed practice (i.e., most participants have lots of practice, but some have little); and 4) bimodal practice (i.e., most participants have either little practice or lots of practice). We then investigated the effects of these different distributions of practice amounts on the validity of the attention control tests.

Goal Reminders

As an additional between-subjects manipulation, we provided on-screen goal reminders (i.e., task instructions) to half of subjects throughout the attention control training sessions in Study 1. One reason subjects are hypothesized to err on conflict tasks such as the Stroop is because they struggle to maintain task goals in an active state (Kane & Engle, 2003). Goal neglect leads to providing the prepotent response on the Stroop task (i.e., reading the word instead of identifying its color), which is an incorrect response on incongruent trials. It has been suggested that individual differences in goal maintenance help explain the relation between working memory capacity and performance on the Stroop paradigm (Kane & Engle, 2003; Hood et al., 2021; Hood et al., 2023). When the proportion of congruent trials is high, low working memory capacity performers commit more errors than those with high working memory capacity, and this is hypothesized to be a result of goal neglect induced by the predominance of non-conflict trials (Kane & Engle, 2003). By contrast, when the proportion of congruent trials is

low, most trials reinforce task goals, and under these conditions working memory capacity predicts response-time interference (Kane & Engle, 2003). Thus, performance on the Stroop task seems jointly determined by goal maintenance and competition resolution, and different task sets (i.e., the proportion of congruent to incongruent trials) can ramp up or down the relative contribution of each of these processes.

Supporting this line of reasoning, Hood and colleagues (Hood et al., 2021) found that goal reminders eliminated the relation between working memory capacity and Stroop errors, when the Stroop task had a large proportion of congruent trials. Their goal reminder manipulation had subjects stop and vocalize task instructions every 12 trials. Providing goal reminders in this manner served to equalize lower- and higher-working memory capacity participants, rendering differences in working memory capacity unrelated to Stroop error rates.

We explored whether goal reminders would exert a similar effect in more complex tasks. The attention control tasks used herein demand goal maintenance and competition resolution; on any given trial the subject must maintain multiple goals and navigate multiple levels of conflict. If goal maintenance is a critical component of these tasks, goal reminders should increase overall performance and reduce the relation between working memory capacity and task performance by offloading memory demands to the task environment. On the other hand, if between-subjects variance in these tasks primarily reflects conflict resolution, then goal reminders ought to have little effect on performance or the validity of the measures. We also explored whether these predictions were altered by practice; goal reminders might exert influence early on when the task is novel and goal-maintenance is paramount (Duncan et al., 1996), but dissipate as subjects gain extensive experience on the tasks.

Study 1 Method

Participants

Participants were recruited from the Georgia Tech and Clemson University undergraduate research pools. They were awarded course credit for participating in the study. Participants were required to be between the ages of 18-35 with normal or corrected-to-normal vision and had to have learned English before age five. The study was IRB approved (protocols H23109 and IRB2023-0187). We recruited participants for four semesters, aiming to amass as large of a sample as possible across both universities.

Procedure

Participants signed up for the study on SONA and were directed to Qualtrics to provide informed consent. Next, they were assigned a subject number and taken to the study dashboard website, where they completed the cognitive ability tests on their personal computers (see Figure 1). They were instructed to try to finish all the tasks in a day or two, and that they should finish any task they start in the same session.

Participants first completed three complex span tests of working memory capacity: Advanced Operation Span, Advanced Rotation Span, and Advanced Symmetry Span (around 45 minutes of testing). Next, they completed three tests of fluid intelligence: Raven's Advanced Progressive Matrices, Number Series, and Letter Sets (around 25 minutes). Finally, they practiced the attention control tests (approximately 45 minutes to 1 hour).

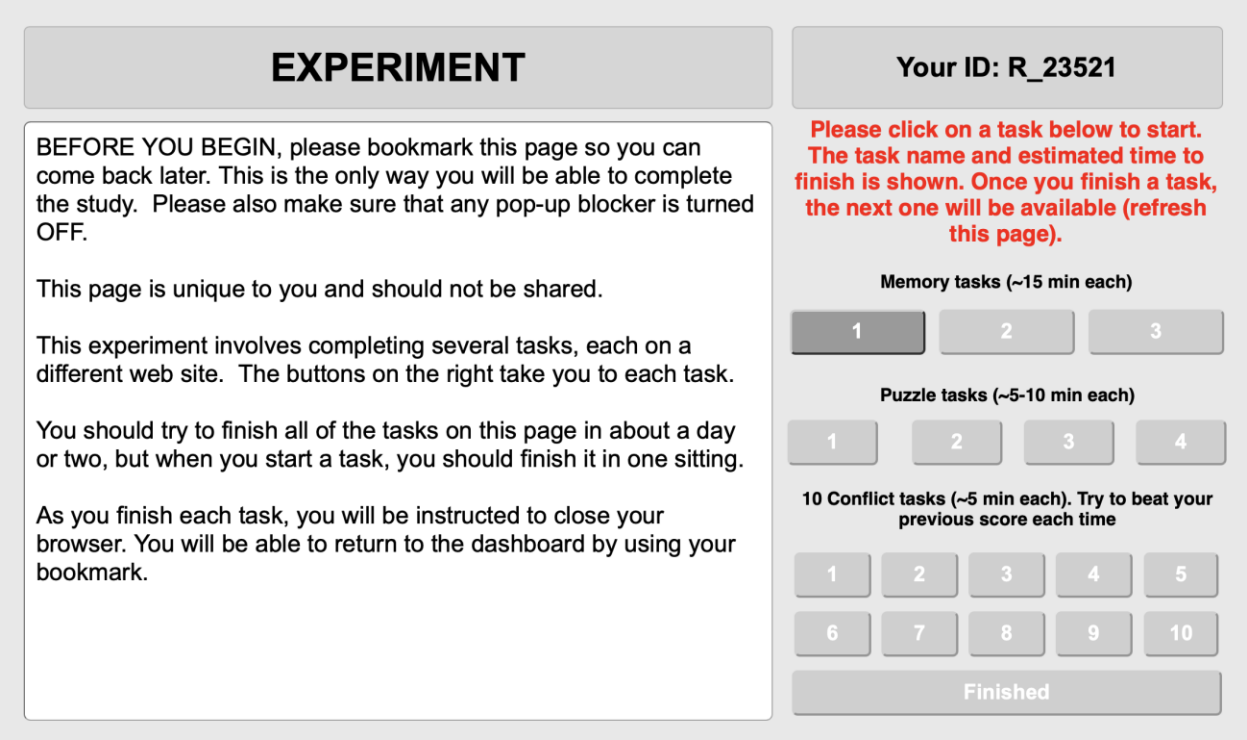
Participants were randomly assigned to either practice the Squared attention control tests or a new attention control test called "Conflict Cubed". Participants in the Squared condition practiced the three Squared attention control tests 10 times. Specifically, they completed 90s of test trials on Stroop Squared, then Flanker Squared, and then Simon Squared, and this constituted

one “round” of training. There were 10 rounds of training on the Squared tasks. The order of the Squared tasks was consistent for all subjects. Participants in the Conflict Cubed condition practiced the 90s Conflict Cubed attention control test 30 times.

As an additional between-subjects manipulation, participants were randomly assigned to either have or not have goal reminders (i.e., task instructions) displayed on the screen for the duration of the attention control tests (see below for more detail).

Figure 1

Depiction of the study dashboard website



Attention Control

For all of the attention control tasks, on the first attempt, participants were shown an example item and were given instructions on how to complete the task. The correct response to the example item was indicated by a green checkmark and a description of why each response option was correct or incorrect. After reading the instructions, participants began a 30-second

practice phase with feedback on every trial. The participant's score and the task timer were displayed at the top of the screen. After 30 seconds of practice, participants were shown their final score on the practice phase and taken back to the instructions screen for further review. After reviewing the instructions again, the participant proceeded to the test phase by clicking the "start" button. A three-second timer counted down, and then the test phase began. Participants were given 90 seconds to earn as many points as possible, earning one point for each correct response and losing one point for each incorrect response.

Feedback was given on every trial in the same manner as during the practice phase. The participant's score and the task timer were displayed at the top of the screen. After the 90s test phase was completed, participants were told their final score. On each attempt following the first one, participants completed 5 seconds of practice before continuing to the 90s test phase.

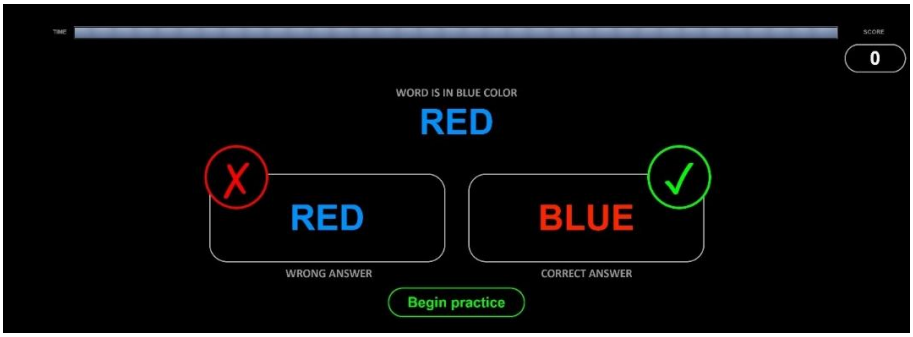
During testing, the mouse cursor was not repositioned between trials; participants moved the cursor from their previous response to the next one. This methodological decision was based on the observation that in most computer applications, cursors do not move without user intervention. Furthermore, the response options were positioned relatively close together on the screen, minimizing the amount of cursor travel required to execute a response.

Stroop Squared (Burgoyne et al., 2023; Figure 2). In Stroop Squared, participants must select the response option with the word meaning that matches the display color of the target stimulus. For example, if the target stimulus is the word "RED" appearing with a blue display color, the participant should select the response option that says the word "BLUE," regardless of the response option's display color. There were four trial types that were sampled randomly: *fully congruent*, *fully incongruent*, *stimulus congruent response options incongruent*, and *stimulus incongruent response options congruent* (see Burgoyne et al. [2023] for additional

details). Participants who received goal reminders were shown text at the top of the screen that said: “Match the COLOR of the top word to the MEANING of the words below.” Scores were computed as the number of correct responses minus the number of incorrect responses.

Figure 2

Stroop Squared



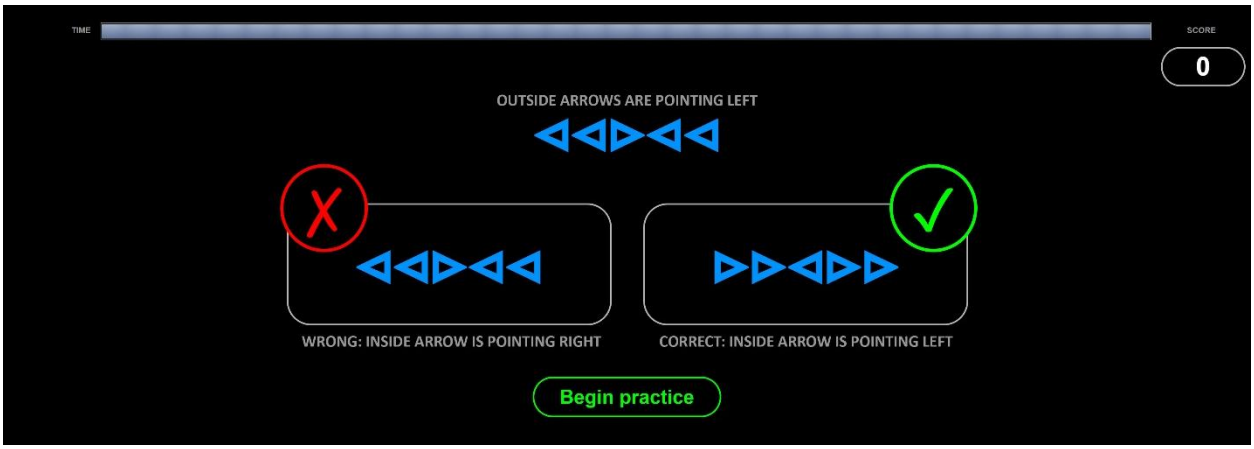
Note. The participant’s task is to select the response option with the word meaning that matches the color of the target stimulus. In the figure, the target stimulus is the word “RED” appearing in blue color, so the participant should select the response option that says the word “BLUE”.

Flanker Squared (Burgoyne et al., 2023; Figure 3). In Flanker Squared, participants must match the direction of the flanking arrows of the target stimulus with the direction of the central arrow of one of two response options. For example, if the target stimulus is “< < > < <”, the participant should select the response option “> > < > >”. The flanking arrows in the target stimulus are pointing to the left, so the participant should select the response option that has a central arrow pointing to the left. There were four trial types that were sampled randomly: *fully congruent*, *fully incongruent*, *stimulus congruent response options incongruent*, and *stimulus incongruent response options congruent* (see Burgoyne et al. [2023] for additional details). Participants who received goal reminders were shown text at the top of the screen that said: “Match the DIRECTION OF THE OUTER ARROWS on top to the DIRECTION OF THE

INNER ARROWS on bottom." Scores were computed as the number of correct responses minus the number of incorrect responses.

Figure 3

Flanker Squared



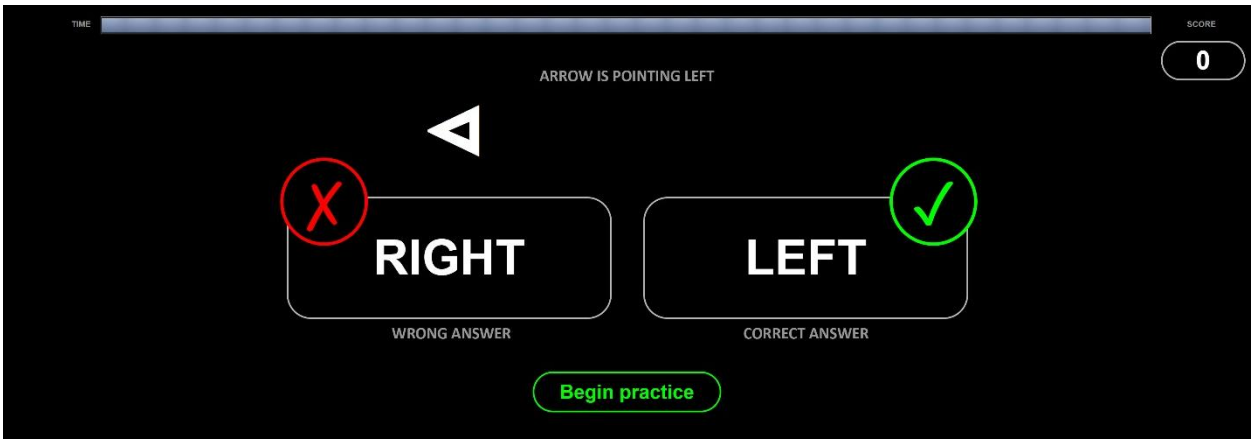
Note. The participant's task is to select the response option with a central arrow that points in the same direction as the flanking arrows in the target stimulus. In the above example, the target stimulus has flanking arrows pointing left, so the participant must select the response option that has a central arrow pointing left.

Simon Squared (Burgoyne et al., 2023; Figure 4). In Simon Squared, participants must match the direction that a target stimulus arrow is pointing with the semantic meaning of one of two response options. For example, if the target stimulus is an arrow pointing to the left, then regardless of which side of the screen it appears on, the participant should select the response option that says the word "LEFT". There were four trial types that were sampled randomly: *fully congruent*, *fully incongruent*, *stimulus congruent response options incongruent*, and *stimulus incongruent response options congruent* (see Burgoyne et al. [2023] for additional details). Participants who received goal reminders were shown text at the top of the screen that said:

"Match the DIRECTION of the arrow to the MEANING of the words below." Scores were computed as the number of correct responses minus the number of incorrect responses.

Figure 4

Simon Squared



Note. The participant's task is to select the response option that states the direction that the arrow is pointing. In the above example, the arrow is pointing left, so the participant must select the response option that says "LEFT".

Conflict Cubed. In Conflict Cubed, participants were presented trials from each of the three Squared tasks, sampled randomly so that any given trial had an equal likelihood of being from Stroop Squared, Flanker Squared, or Simon Squared. As an introduction, subjects were given instructions and 30 seconds of practice on Stroop Squared, Flanker Squared, and Simon Squared, one after another. During the 90 second test phases of the Cubed task, trials from the three Squared tasks were randomly sampled. Participants who received goal reminders were shown the same text as participants who were assigned to practice the Squared tasks, with the goal reminder updating on every trial according to whether the trial was taken from Stroop

Squared, Flanker Squared, or Simon Squared. Scores were computed as the number of correct responses minus the number of incorrect responses.

Fluid Intelligence

Raven's Advanced Progressive Matrices (Raven & Court, 1998). In Raven's Matrices, participants were shown a 3x3 grid of abstract patterns with the pattern in the bottom right corner missing. The participant's task was to select from 8 response options the pattern that best fit the grid. We gave participants 10 minutes for 18 items from Raven's Advanced Progressive Matrices; the measure of performance was the number of items they correctly responded to.

Letter Sets (Ekstrom et al., 1976). In letter sets, participants were shown five sets of four letters and challenged to identify the set of letters that did not follow the same pattern as the others. We gave participants 10 minutes to complete 20 items; the measure of performance was the number of items they correctly responded to.

Number Series (Thurstone, 1938). In number series, we presented participants with a set of numbers that followed a pattern. They were shown five possible response options that could complete the pattern and asked to select the response option that best followed the pattern of the number series. We gave participants 5 minutes for 15 items; the measure of performance was the number of items they correctly responded to.

Working Memory Capacity

Advanced Symmetry Span (Kane et al., 2004; Unsworth et al., 2005; Draheim et al., 2018). In symmetry span, participants must remember spatial locations while deciding whether patterns are symmetrical or not. On a given trial, the participant was shown a symmetrical or asymmetrical grid and needed to determine whether or not it was symmetrical. Next, they were shown a 4x4 grid of squares, and one of them was emphasized by a red color. Their goal was to

memorize the location of the colored square. This symmetry/square interleaving pattern continued for 2 to 7 times (i.e., the set sizes used in the task). Afterward, the participant needed to report the location that the colored squares appeared in, in the order that they appeared. We gave participants 12 trials: 2 of each set size. We used the partial scoring method, awarding one point for each memory item that was correctly recalled (Conway et al., 2005).

Advanced Rotation Span (Kane et al., 2004; Unsworth et al., 2005; Draheim et al., 2018). In rotation span, participants remembered directional arrows while deciding whether a letter was in the proper orientation or mirror-imaged orientation. On a given trial, the participant was shown a letter they would mentally rotate to determine its orientation (mirror-imaged or normal). Next, they were shown a single arrow that was either small or large and pointed in one of 8 directions. This letter/arrow interleaving pattern continued 2 to 7 times (i.e., the set sizes used in the task). Afterward, the participant was asked to report the arrows in the order they appeared. We gave participants 12 trials: 2 of each set size. We used the partial scoring method, awarding one point for each memory item that was correctly recalled (Conway et al., 2005).

Advanced Operation Span (Kane et al., 2004; Unsworth et al., 2005; Draheim et al., 2018). In operation span, participants remembered letters while deciding whether simple arithmetic equations were accurate or inaccurate. On a given trial, the participant was shown a simple arithmetic equation they would solve to determine whether it was accurate or inaccurate. Next, they were shown a single letter to remember. This math/letter interleaving pattern continued 3 to 9 times (i.e., the set sizes used in the task). Afterward, the participant was asked to report the letters in the order they appeared. We gave participants 14 trials: 2 of each set size. We used the partial scoring method, awarding one point for each memory item that was correctly recalled (Conway et al., 2005).

Transparency and Openness

We describe our data exclusions, all manipulations, and all measures in the study. Data and analysis code for Study 1 are available at https://osf.io/d5bzy/. Data were analyzed using R, version 4.0.0 (R Core Team, 2020), and IBM SPSS Statistics (version 27).

Data Preparation

We removed participants' scores on the cognitive ability tests if they showed severely poor performance indicating they did not understand the instructions or were not performing the task as intended. For the three fluid intelligence tests, we removed scores that were at or below chance-level performance. For the three working memory tests, we removed scores if participants completed the distractor task with an accuracy rate of 60% or worse (for reference, chance-level performance would be 50%). Additionally, for the fluid intelligence and working memory tests, we removed outlying scores, which were defined as standardized scores (i.e., Z-scores) worse than 3.5 *SD*s from the sample mean. There were no outlying scores; see the Supplemental Materials for more information.

For the attention control tests, we removed scores that were at or below 5 points, as these scores were likely negatively affected by lack of effort or misunderstanding of the task. On any attempt in which a participant earned 5 points or less, we also set the associated accuracy, response time, and trial count data to missing. We also removed scores (as well as associated accuracy, response time, and trial count data) on any performance attempt in which the mean response time was greater than or equal to 8 seconds per trial, reasoning that these performance attempts may have been negatively affected by internet problems or off-task distractions. A detailed breakdown of the data processing pipeline and effective *n* for each measure at each stage of processing is provided in the Supplemental Materials.

Study 1 Results

Combined, our sample includes 566 participants contributing to the primary analyses of interest on the attention control tests. Due to data preparation procedures (see Supplemental Materials), the final *n* differs across measures, as shown in Table 1. Our analytical sample includes 359 participants with at least one valid score on one of the Squared tasks; of these participants, 207 received goal reminders and 152 did not. The sample also includes 216 participants with at least one valid score on the Conflict Cubed task; of these participants, 108 received goal reminders and 108 did not.

Descriptive statistics for the cognitive ability measures are presented in Table 1. The four attention control tests had very high estimates of internal consistency reliability on the first attempt (split-half reliability = .99) and on the final attempt (split-half reliability = .99), which is likely a consequence of their speeded, points-based format. Internal consistency was adequate for the three fluid intelligence tests (.66, .67, and .81) and high for the three complex span tests of working memory capacity (ranging from .83 to .84).

Between-subjects variability increased with practice. On Stroop Squared, the *SD* increased from 10.78 to 13.35 from the first to the last attempt, and the correlation between the *SD* and the attempt number was positive and statistically significant ($r = .86$, $p = .002$). On Flanker Squared, the *SD* increased from 10.42 to 13.27 and the correlation with attempt number was statistically significant ($r = .88$, $p < .001$). On Simon Squared, the *SD* increased from 10.63 to 12.91 and the correlation with attempt number was statistically significant ($r = .89$, $p < .001$). On Conflict Cubed, the *SD* increased from 10.20 to 11.24 and the correlation with attempt number was statistically significant ($r = .69$, $p < .001$). Thus, there was no evidence of restriction

of range after subjects had engaged in extensive practice; rather, there was evidence of range *enhancement*.

Correlations are presented in Table 2. The three Squared tests of attention control were robustly correlated with each other on the first attempt (*r*s of .39, .36, and .40), and these correlations increased after extensive practice (final attempt *r*s of .59, .56, and .57). The three measures of fluid intelligence were moderately to highly correlated with each other (*r*s of .35, .42, and .45), as were the three measures of working memory capacity (*r*s of .42, .44, and .69).

Table 1

Descriptive statistics for the cognitive ability measures

Measure	***N***	***M***	***SD***	**Skew**	**Kurtosis**	**Reliability**
Raven's Matrices	531	9.60	3.33	-0.24	-0.72	.81
Letter Sets	531	9.31	2.73	0.36	-0.54	.66
Number Series	546	9.65	2.70	-0.25	-0.82	.67
Symmetry Span	568	32.40	12.78	-0.59	-0.30	.84
Operation Span	567	58.35	16.43	-0.97	1.02	.83
Rotation Span	562	29.56	11.62	-0.37	-0.47	.83
Conflict Cubed 1st Attempt	197	25.26	10.20	-0.14	-0.81	.99
Conflict Cubed 30th Attempt	162	38.59	11.23	-0.54	-0.10	.99
Stroop Squared 1st Attempt	329	29.88	10.78	-0.39	-0.45	.99
Stroop Squared 10th Attempt	317	39.65	13.35	-0.49	-0.31	.99
Flanker Squared 1st Attempt	300	27.95	10.42	-0.12	-0.08	.99
Flanker Squared 10th Attempt	300	37.71	13.27	-0.16	-0.65	.99
Simon Squared 1st Attempt	331	48.52	10.63	-1.19	2.16	.99
Simon Squared 10th Attempt	304	51.19	12.91	-1.15	1.30	.99

Note. Reliability refers to split-half reliability with a Spearman-Brown correction.

Table 2

Correlations among the cognitive ability measures

Measure	Raven's Matrices	Letter Sets	Number Series	Symmetry Span	Operation Span	Rotation Span	Conflict Cubed 1	Conflict Cubed 30	Stroop Squared 1	Stroop Squared 10	Flanker Squared 1	Flanker Squared 10	Simon Squared 1	Simon Squared 10
Raven's Matrices	---													
Letter Sets	**.35**	---												
Number Series	**.45**	**.42**	---											
Symmetry Span	**.42**	**.35**	**.34**	---										
Operation Span	**.29**	**.29**	**.22**	**.42**	---									
Rotation Span	**.37**	**.27**	**.26**	**.69**	**.44**	---								
Conflict Cubed 1	**.41**	**.24**	**.28**	**.34**	**.23**	**.33**	---							
Conflict Cubed 30	**.33**	**.33**	**.32**	**.31**	**.20**	**.27**	**.43**	---						
Stroop Squared 1	**.29**	**.28**	**.27**	**.36**	**.19**	**.32**	---	---	---					
Stroop Squared 10	**.22**	**.18**	**.27**	**.30**	**.25**	**.26**	---	---	**.39**	---				
Flanker Squared 1	**.38**	**.33**	**.37**	**.37**	**.24**	**.29**	---	---	**.39**	**.34**	---			
Flanker Squared 10	**.32**	**.25**	**.24**	**.33**	**.19**	**.32**	---	---	**.25**	**.59**	**.51**	---		
Simon Squared 1	**.26**	**.29**	**.33**	**.31**	**.20**	**.31**	---	---	**.36**	**.39**	**.40**	**.31**	---	
Simon Squared 10	**.25**	**.17**	**.28**	**.24**	**.21**	**.24**	---	---	**.20**	**.56**	**.28**	**.57**	**.52**	---

Note. Pairwise *n* ranges from 150 to 563. The digit after the attention control tests refers to the attempt number. Correlations between Conflict Cubed and the Squared tasks could not be estimated because participants were assigned to complete one set of tasks or the other. All correlations were significant $p < .05$.

Number of Trials

We computed the number of trials completed by participants over repeated testing to facilitate comparisons to other studies of practice effects. Specifically, we summed the number of trials completed across all practice attempts, after filtering the data to only include participants who completed all of the assigned practice sessions. On Conflict Cubed, participants completed an average of 1249.20 total trials (*SD* = 196.64, range = 634 to 1671). On Stroop Squared, they completed an average of 453.34 trials (*SD* = 77.74, range = 249 to 624). On Flanker Squared, they completed an average of 405.51 trials (*SD* = 82.09, range = 197 to 623). On Simon Squared, they completed an average of 561.08 trials (*SD* = 76.76, range = 297 to 725). Detailed descriptive statistics are provided in the Supplemental Materials.

On each attempt, participants completed an average of 40.16 trials on Conflict Cubed (*SD* = 10.43, range = 8 to 182). On Stroop Squared, they completed an average of 42.99 trials per attempt (*SD* = 11.98, range = 6 to 194). On Flanker Squared, they completed an average of 39.41 trials per attempt (*SD* = 11.69, range = 6 to 129). On Simon Squared, they completed an average of 52.21 trials per attempt (*SD* = 12.34, range = 6 to 132).

Psychometrics of the Attention Control Tests

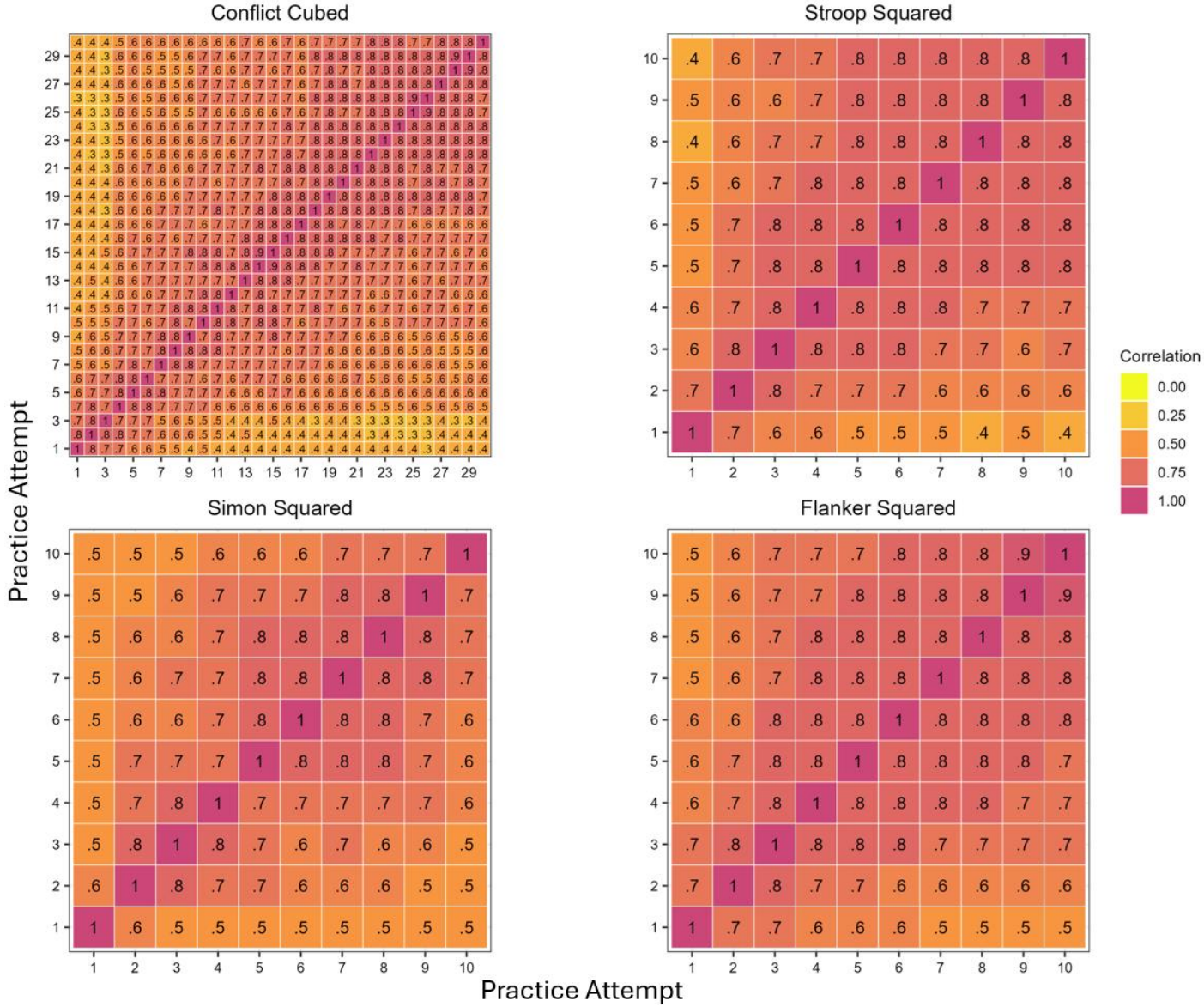
We estimated the test-retest reliability of the attention control tests over repeated administrations and found a simplex structure in which proximal attempts were more closely related (see Figure 5 for a test-retest reliability heatmap). From the first attempt to the second attempt, test-retest reliability was *r* = .66, .74, and .62 for Stroop Squared, Flanker Squared, and Simon Squared, respectively, and *r* = .81 for Conflict Cubed. Generally, test-retest reliability decreased from the first attempt to subsequent attempts. As one might expect, extensive practice on the tasks led to lower retest reliability when comparing performance on the first and last

attempts: r = .39, .51, .52 for Stroop Squared, Flanker Squared, and Simon Squared, respectively, and r = .43 for Conflict Cubed.

Interestingly, for Stroop Squared and Simon Squared, retest reliability increased significantly when examining the correlation between performance on the *second* and *third* attempt (i.e., Stroop Squared increased from .66 to .79, $Z = 4.25$ $p < .001$; Simon Squared increased from .62 to .76, $Z = 3.95$, $p < .001$). Thus, for these tasks, the stability of participants' rank order improved following the first attempt. We return to this point in the Discussion, as it suggests that providing participants with slightly more practice before the testing phase could increase rank-order stability.

Figure 5

Correlation heatmap depicting test-retest reliability across all attempts on the attention control tasks



Performance on the Attention Control Tasks with Practice and Goal Reminders

Figures illustrating performance as a function of practice and goal reminders are presented in Figure 6a; figures depicting accuracy rates and response times are presented in Figures 6b and 6c.

We tested a series of repeated measures mixed models to estimate the effects of practice, goal reminders, fluid intelligence, and working memory capacity on performance in the attention control tests. We used the "GENLINMIXED" command in SPSS to estimate generalized linear models, which use maximum likelihood estimation with robust standard errors for fixed effects and allowed us to retain participants with partially missing data. Observations were nested within subjects, and we specified an autoregressive covariance structure to account for correlated errors over time. We specified two random effects—a random intercept and a random slope for practice—to allow individual differences in starting point and rate of change with practice; these random effects were allowed to correlate. The models included attempt number and goal reminder condition as fixed effects, as well as two factor scores representing fluid intelligence and working memory capacity. These factor scores were formed using the three indicator measures for each construct and were estimated separately using the 'umx' package in R (Bates, Neale, & Maes, 2019).[1]

We used a model building strategy to compare the fit of four candidate models for each task. Specifically, we tested whether model fit improved by including a random slope for practice, which it did. We also tested whether model fit improved by including three-way interaction terms (between practice, goal reminder condition, and cognitive ability). The three-

[1] Factor loadings for fluid intelligence: Raven's Matrices (.62), Letter Sets (.59), Number Series (.74). Factor loadings for working memory capacity: Symmetry Span (.81), Operation Span (.52), Rotation Span (.87). Factor scores across the two constructs were correlated, $r = .46$, $p < .001$, $n = 576$.

way interaction terms had negligible effects on model fit, but we retained them due to their theoretical relevance (e.g., the effect of goal reminders may be particularly important early in practice and for lower-ability participants; Duncan et al., 1996). A model comparison table and additional results are presented in the Supplemental Materials.

The results of the generalized linear models are presented in Tables 3 and 4. We report both omnibus fixed effects (Type III tests) and individual fixed coefficients. The fixed effects table uses *F*-tests to evaluate whether predictors contribute to the overall model, whereas the fixed coefficients table provides specific parameter estimates with *t*-tests. These may differ when a predictor's overall contribution is significant, but individual contrasts are not, or vice versa. We interpret the omnibus fixed effects to determine predictor significance.

The models revealed that participants significantly improved their scores on Conflict Cubed, Stroop Squared, and Flanker Squared with practice, but not Simon Squared. Goal reminders did not have a significant effect on performance, nor did they interact with any other predictor. Thus, in contrast to our hypothesis, goal reminders did not reduce the role of working memory capacity, nor did they have a stronger effect early during practice. On the other hand, fluid intelligence and working memory capacity had significant main effects on all four tasks, indicating that participants with greater cognitive ability performed better than their lower-ability peers. Surprisingly, for most tasks, there was no significant interaction between practice and the cognitive ability factors, indicating that the relationship between cognitive ability and performance did not change with practice. The exception was Flanker Squared, which revealed a significant interaction between practice and fluid intelligence, as well as practice and working memory capacity.

Figure 6a

Performance on the attention control tasks as a function of practice and goal reminders.
Error bars represent 95% CI, calculated separately at each time point.

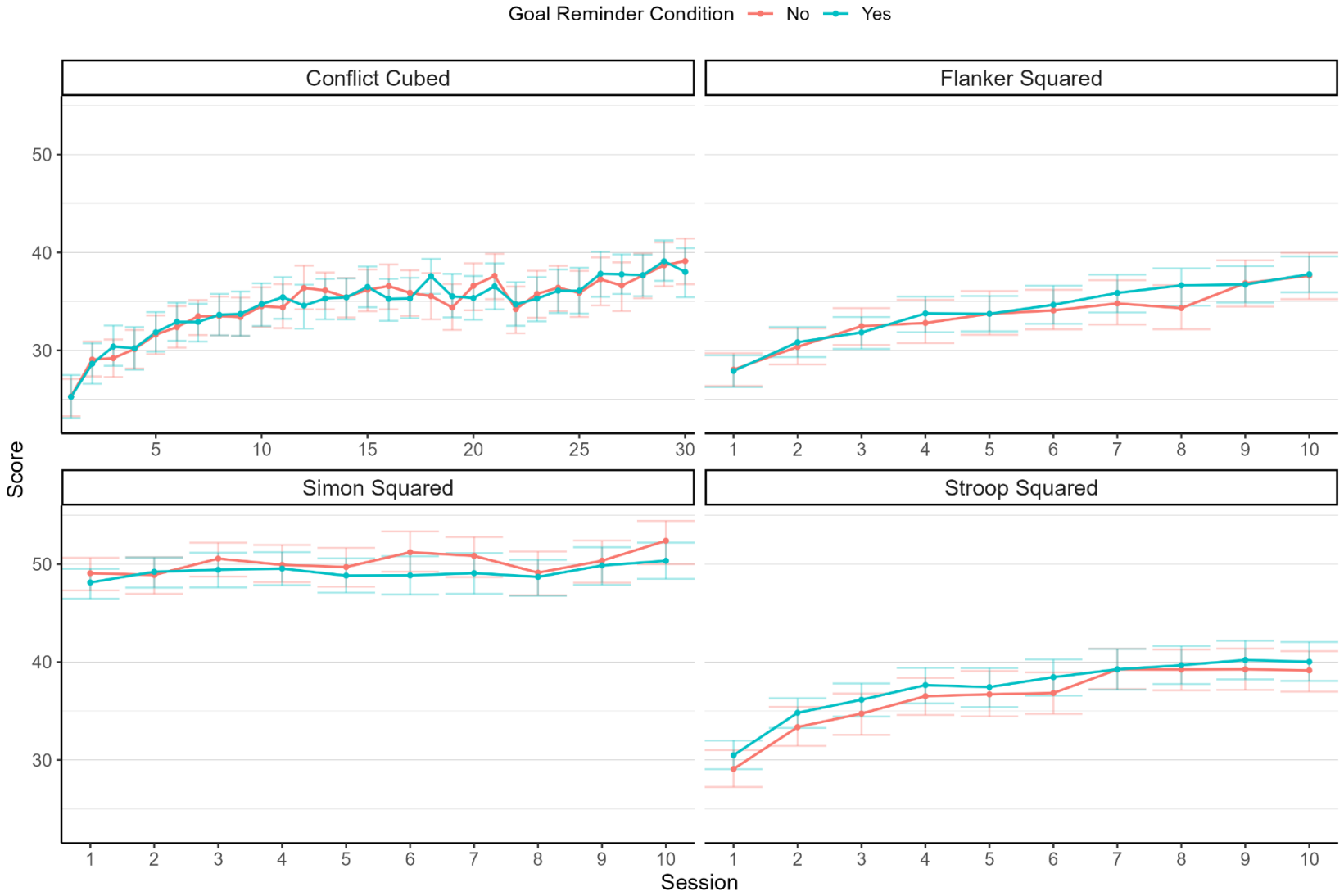


Figure 6b

Response time on the attention control tasks as a function of practice and goal reminders.
Error bars represent 95% CI, calculated separately at each time point.

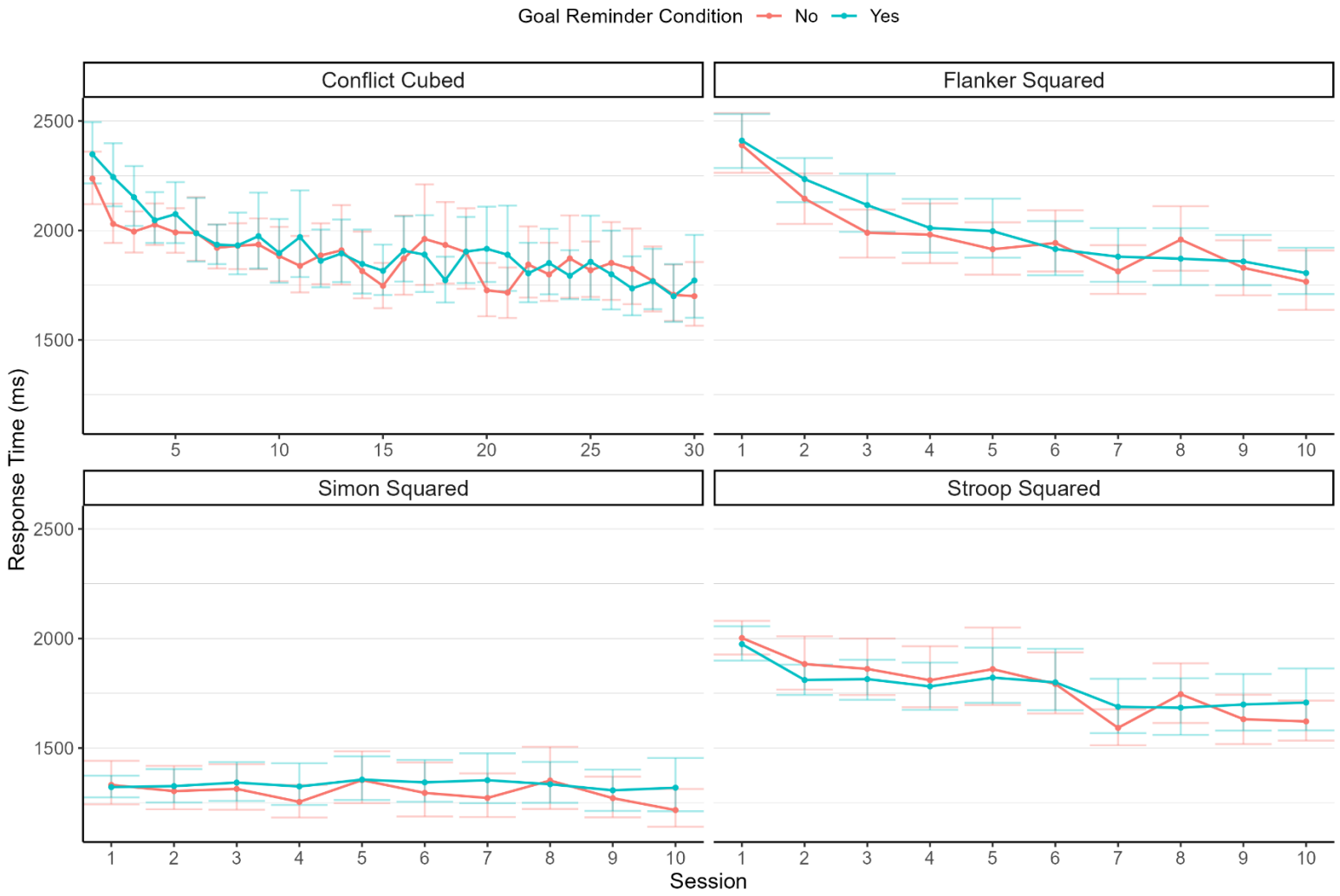


Figure 6c

Accuracy on the attention control tasks as a function of practice and goal reminders.
Error bars represent 95% CI, calculated separately at each time point.

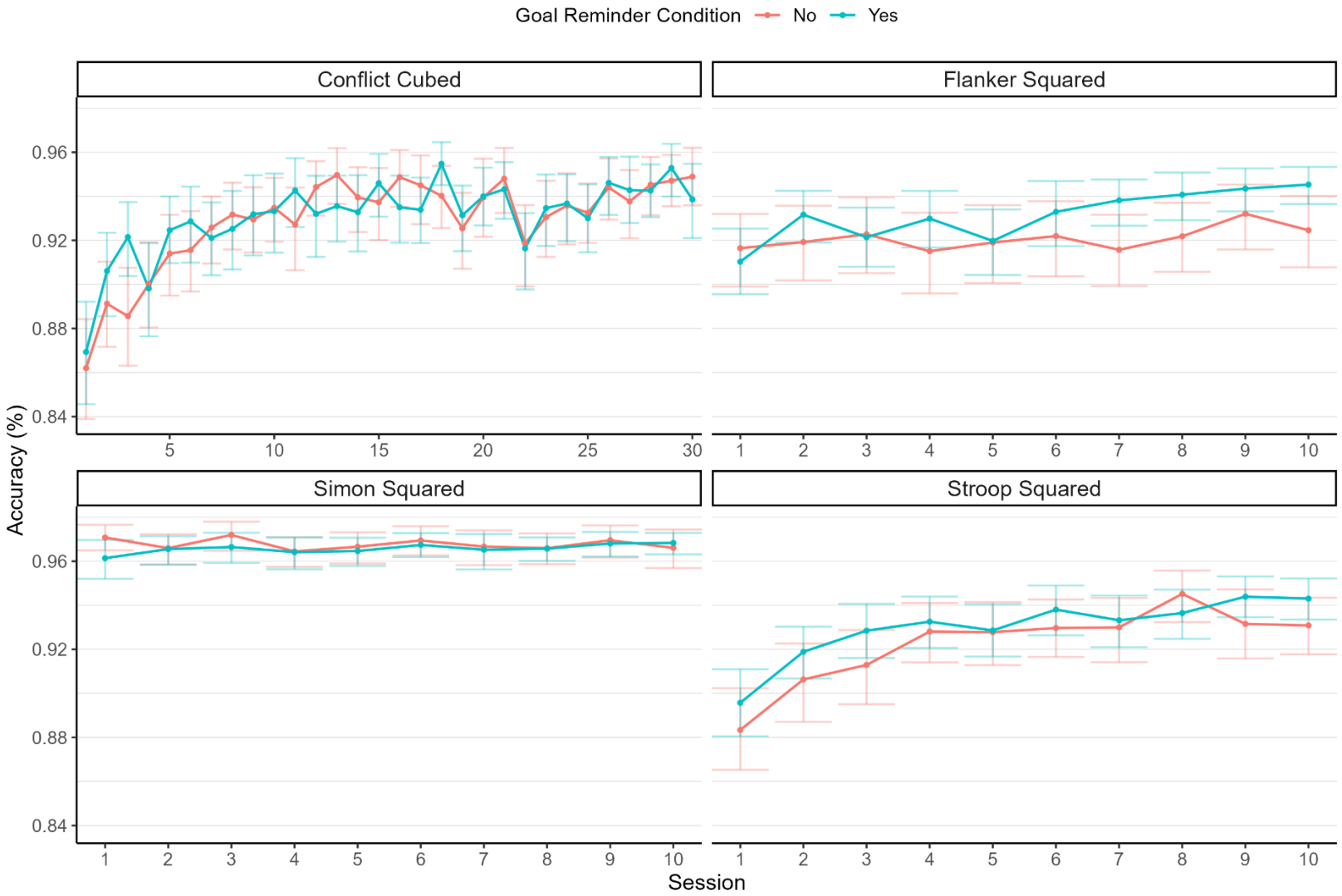


Table 3

Results of Generalized Linear Models Predicting Attention Control Scores: Fixed Effects

	Conflict Cubed		**Stroop Squared**		**Flanker Squared**		**Simon Squared**	
Predictor	***F***	***p***	***F***	***p***	***F***	***p***	***F***	***p***
Practice	**148.61**	**<.001**	**165.24**	**<.001**	**203.06**	**<.001**	2.24	.134
Goal Reminder Condition	0.02	.888	1.37	.242	0.03	.863	0.09	.766
Fluid Intelligence	**35.68**	**<.001**	**36.19**	**<.001**	**77.67**	**<.001**	**29.85**	**<.001**
Working Memory Capacity	**27.92**	**<.001**	**19.27**	**<.001**	**23.03**	**<.001**	**7.38**	**.007**
Practice × Goal Reminders	0.01	.926	0.24	.626	0.01	.905	2.12	.145
Practice × Gf	3.45	.063	0.12	.726	**4.28**	**.039**	0.00	.965
Practice × WMC	3.50	.062	1.33	.250	**4.20**	**.041**	1.36	.243
Goal Reminders × Gf	0.09	.760	0.02	.881	0.29	.589	0.00	.967
Goal Reminders × WMC	0.28	.596	0.04	.841	0.58	.446	0.99	.320
Practice × Goal Reminders × Gf	0.68	.410	1.84	.175	0.24	.625	1.17	.279
Practice × Goal Reminders × WMC	0.15	.697	0.03	.872	2.51	.113	0.30	.584

Note. Gf = fluid intelligence, WMC = working memory capacity. For predictor terms including Goal Reminder condition, values represent the effect when goal reminders were present.

Table 4

Results of Generalized Linear Models Predicting Attention Control Scores: Fixed Coefficients

	Conflict Cubed			**Stroop Squared**			**Flanker Squared**			**Simon Squared**		
Predictor	***b***	***t***	***p***	***b***	***t***	***p***	***b***	***t***	***p***	***b***	***t***	***p***
Practice	**0.31**	**8.05**	**<.001**	**0.97**	**9.21**	**<.001**	**1.00**	**8.90**	**<.001**	**0.21**	**2.02**	**.044**
Goal Reminder Condition	-0.14	-0.14	.888	1.29	1.17	.242	0.17	0.17	.863	0.31	0.30	.766
Fluid Intelligence	**3.90**	**3.97**	**<.001**	**4.26**	**3.83**	**<.001**	**5.98**	**5.22**	**<.001**	**4.04**	**3.47**	**.001**
Working Memory Capacity	**2.76**	**3.55**	**<.001**	**2.99**	**2.60**	**.009**	**2.71**	**2.50**	**.012**	1.23	1.20	.230
Practice × Goal Reminders	0.00	0.09	.926	-0.07	-0.49	.626	0.02	0.12	.905	-0.20	-1.46	.145
Practice × Gf	0.04	0.71	.475	-0.16	-1.19	.233	-0.25	-1.58	.115	-0.10	-0.67	.505
Practice × WMC	-0.08	-1.63	.103	0.09	0.69	.492	**0.32**	**2.31**	**.021**	0.15	1.29	.196
Goal Reminders × Gf	0.42	0.31	.760	0.22	0.15	.881	0.78	0.54	.589	-0.06	-0.04	.967
Goal Reminders × WMC	0.62	0.53	.596	0.29	0.20	.841	1.02	0.76	.446	1.42	0.99	.320
Practice × Goal Reminders × Gf	0.06	0.82	.410	0.26	1.36	.175	0.10	0.49	.625	0.20	1.08	.279
Practice × Goal Reminders × WMC	0.03	0.39	.697	0.03	0.16	.872	-0.28	-1.58	.113	-0.10	-0.55	.584

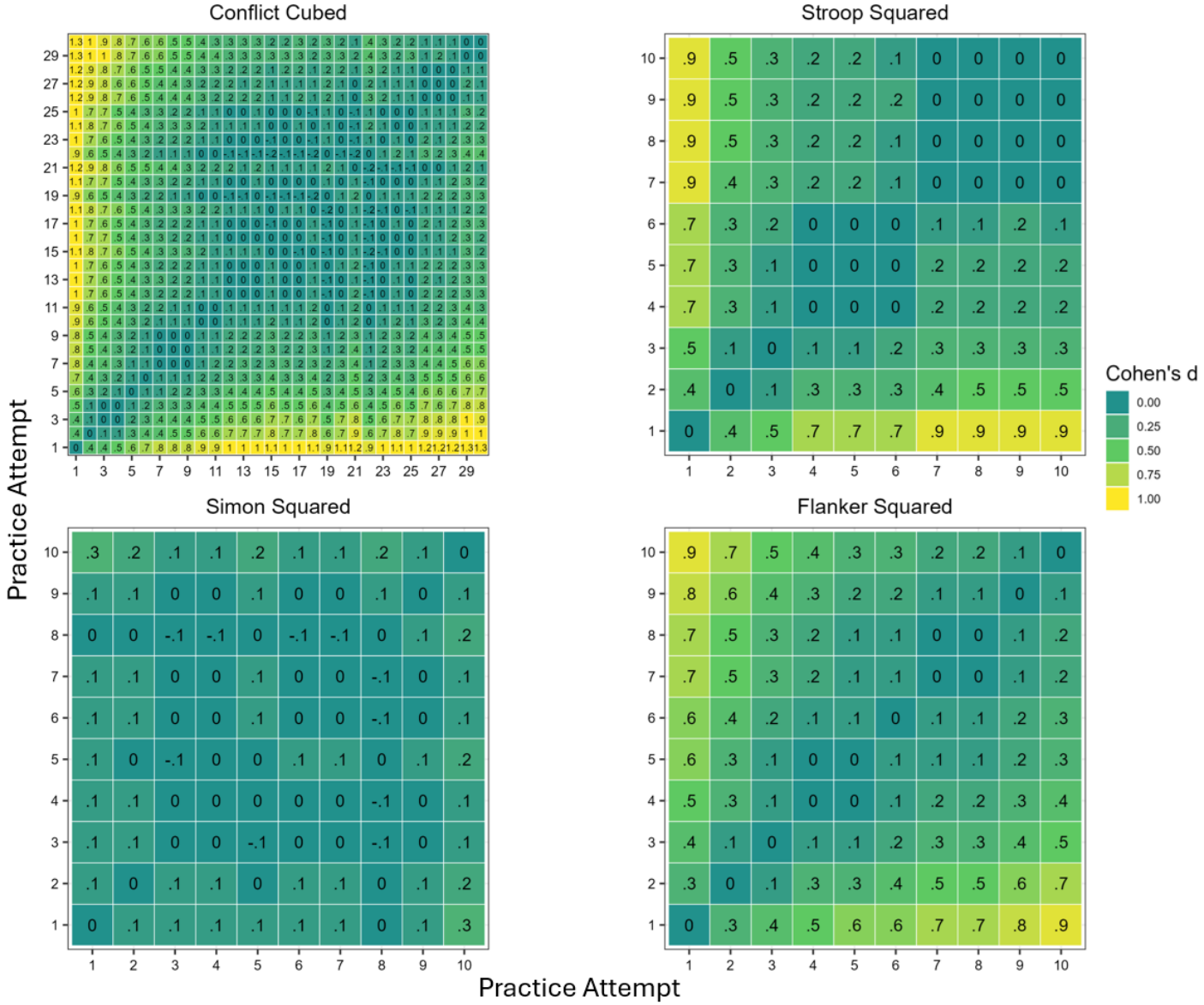
Note. Gf = fluid intelligence, WMC = working memory capacity. For predictor terms including Goal Reminder condition, values represent the effect when goal reminders were present.

Practice Effects on Performance on the Attention Control Tests

We quantified the improvement in performance on the attention control tests due to practice in terms of Cohen's *d*, the standardized mean difference (see Figure 7). For each pairwise comparison, we used the standard deviation of the earlier attempt as the standardizer. Practice effects were much smaller on Simon Squared than the other attention control tests. From the first attempt to the second attempt, the gain in performance was d = .40, .26, and .05 for Stroop Squared, Flanker Squared, and Simon Squared, respectively, and d = .35 for Conflict Cubed. From the first attempt to the third attempt, the gain in performance was slightly larger, d = .53, .40, and .13 for Stroop Squared, Flanker Squared, and Simon Squared, respectively, and d = .44 for Conflict Cubed. From the first attempt to the 10th attempt, which was the last attempt on the Squared tasks, the gain in performance was d = .91, .94 and .25 for Stroop Squared, Flanker Squared, and Simon Squared, respectively, and d = .92 for Conflict Cubed. Finally, for Conflict Cubed, from the first attempt to the 30th attempt, the gain in performance was d = 1.31.

Figure 7

Practice effects on the attention control tests



How Are Correlations with Fluid Intelligence and Working Memory Capacity Affected by Practice?

Next, we examined how the correlation between attention control and the two cognitive ability factors, fluid intelligence and working memory capacity, changed as a function of practice and goal reminders. We illustrated these relationships in Figures 8a and 8b. Collapsing across goal reminder conditions, there was generally a negative trend between the attempt number and the magnitude of the correlation between the cognitive ability factors and performance on the attention control tasks. However, the mixed models we reported previously in Table 3 indicated that the interactions between the cognitive ability factors and practice were not statistically significant, except for Flanker Squared. Here, we present the observed correlations at each time point, along with their sample sizes and 95% confidence intervals, to provide greater transparency on the magnitude of the effects.

For Conflict Cubed, the correlation with fluid intelligence was r = .40, 95% CI [.27, .51], on Attempt 1 and r = .41, 95% CI [.27, .53] on Attempt 10 (Δr = +.01), while the correlation with working memory capacity was r = .37, 95% CI [.24, .48] on Attempt 1 and r = .31, 95% CI [.17, .44] on Attempt 10 (Δr = -.06). The effective n was 197 on the first attempt and 162 on the final attempt.

For Stroop Squared, the correlation with fluid intelligence was r = .35, 95% CI [.26, .44], on Attempt 1 and r = .30, 95% CI [.20, .40] on Attempt 10 (Δr = -.05), while the correlation with working memory capacity was r = .36, 95% CI [.26, .45] on Attempt 1 and r = .32, 95% CI [.21, .41] on Attempt 10 (Δr = -.04). The effective n was 329 on the first attempt and 317 on the final attempt.

For Flanker Squared, the correlation with fluid intelligence was r = .47, 95% CI [.38, .56] on Attempt 1 and r = .34, 95% CI [.24, .44] on Attempt 10 (Δr = -.13), while the correlation with working memory capacity was r = .36, 95% CI [.26, .46] on Attempt 1 and r = .35, 95% CI [.25, .45] on Attempt 10 (Δr = -.01). The effective n was 300 on the first and final attempt.

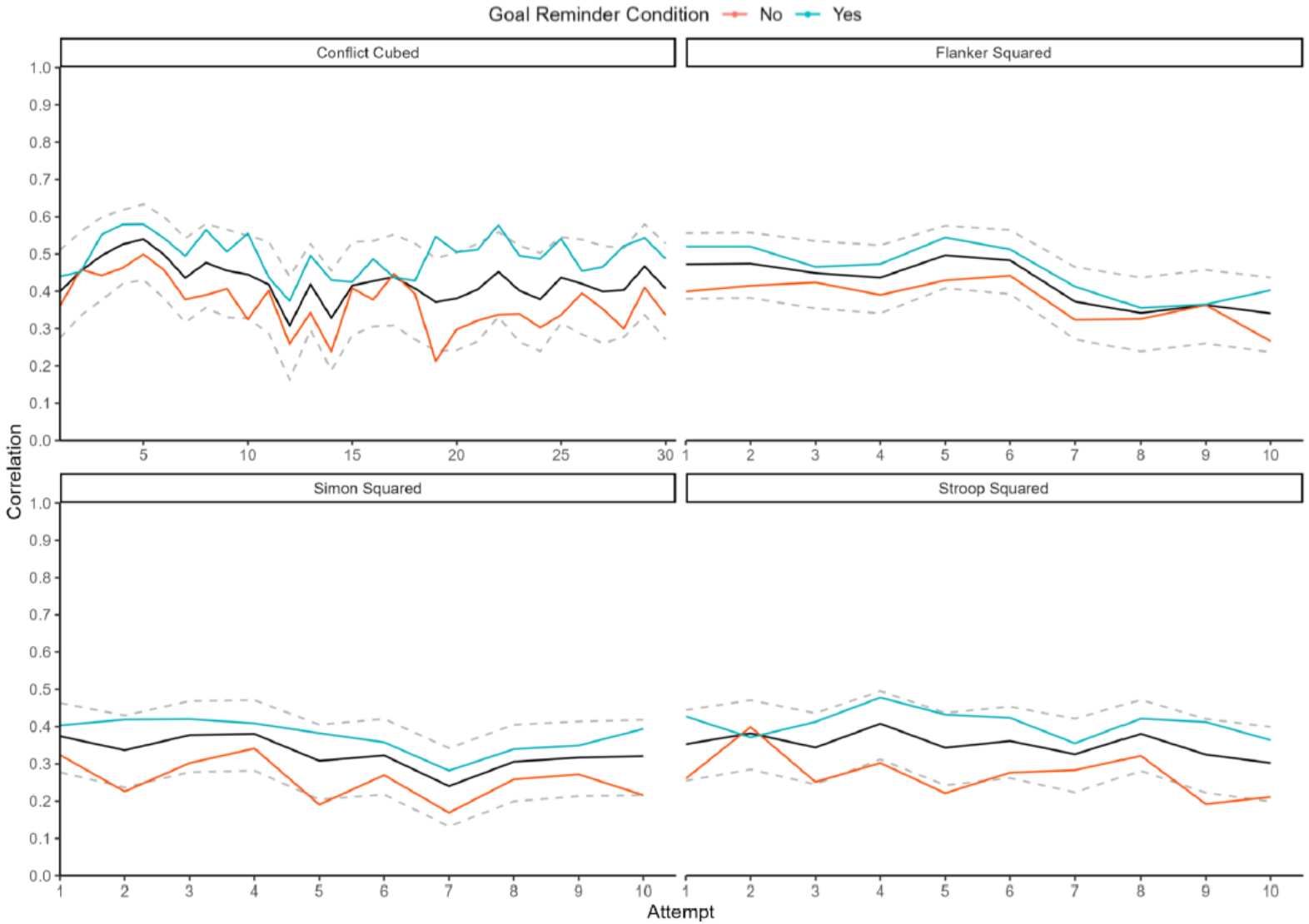
For Simon Squared, the correlation with fluid intelligence was r = .37, 95% CI [.28, .46] on Attempt 1 and r = .32, 95% CI [.22, .42] on Attempt 10 (Δr = -.05), while the correlation with working memory capacity was r = .34, 95% CI [.24, .43] on Attempt 1 and r = .28, 95% CI [.17, .38] on Attempt 10 (Δr = -.06). The effective n was 331 on the first attempt and 304 on the final attempt.

To summarize, for relations with fluid intelligence, the average change in correlation from the first attempt to the last was Δr = -.06, with a range of -.13 to +.01. For relations with working memory capacity, the average change in correlation from the first attempt to the last was Δr = -.04, with a range of -.06 to -.01. Overall, the magnitude of the effect of practice on the relation between the attention control measures and fluid intelligence and working memory was small.

One other interesting finding in Figures 8a and 8b is that the correlations between attention control and the cognitive ability factors were numerically larger for participants in the goal reminder condition, as depicted by the blue curve, relative to participants in the no-goal-reminder condition, depicted by the red curve. Although the mixed models revealed that these interactions were not significant, they trend in the opposite direction of our hypothesis; we had predicted that attention control performance would be more strongly related to cognitive ability when participants had to keep track of task instructions without the help of on-screen goal reminders—this was not the case.

Figure 8a

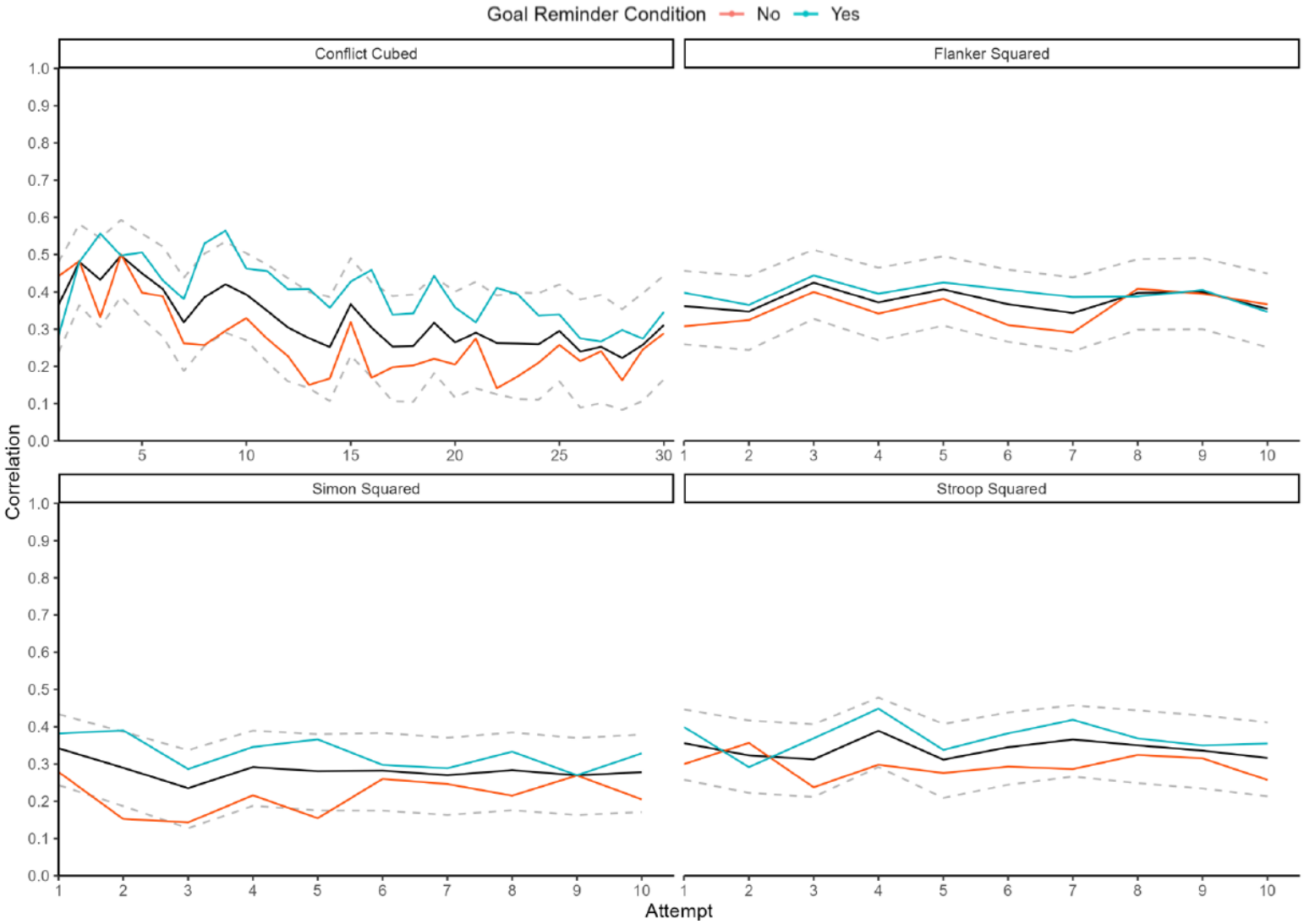
Correlations between attention control and fluid intelligence as a function of practice and goal reminder condition



Note. Black = overall, collapsing across goal reminder condition; Dashed = overall 95% CI.

Figure 8b

Correlations between attention control and working memory capacity by practice and goal reminder condition



Note. Black = overall, collapsing across goal reminder condition; Dashed = overall 95% CI.

What Happens When Different People Have Different Amounts of Practice?

In high-stakes testing scenarios, test takers may have different amounts of practice on selection tests. We simulated how this would affect the correlation between performance on the attention control tests and the fluid intelligence and working memory capacity factor scores. If correlations are considerably reduced when different participants have different amounts of practice, this would indicate that the rank-ordering of subjects is tenuous enough that individual differences in ability can be overcome with practice, which could be problematic from a selection standpoint.

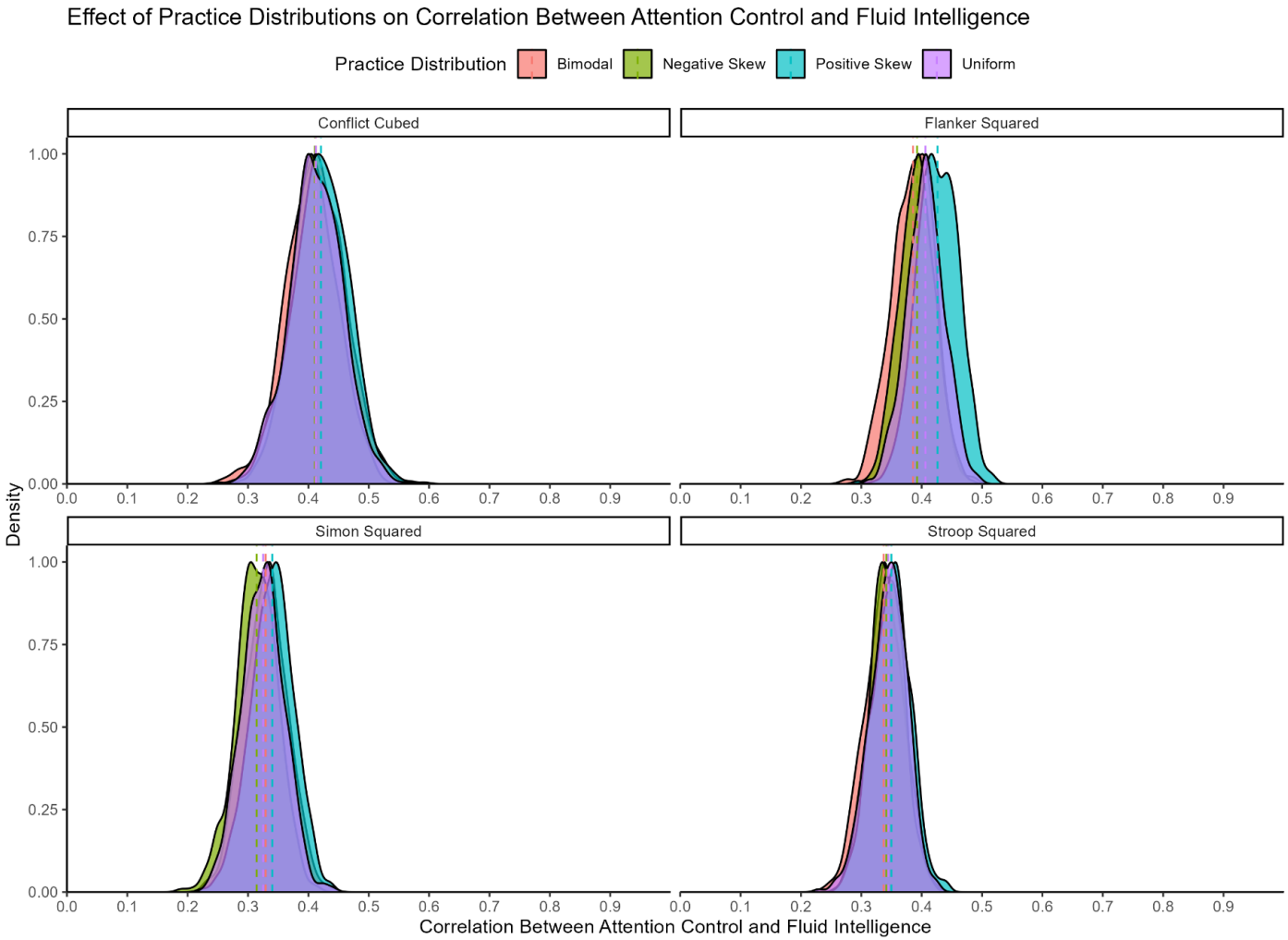
We simulated four different distributions of practice by sampling attention control scores from different practice attempts for each subject, iterating over this sampling procedure 1,000 times for each simulated distribution of practice. We assessed the following distributions of practice within a sample: 1) uniform distribution (i.e., each participant has an equal likelihood to have any amount of practice); 2) positively-skewed practice (i.e., most participants have little practice, but some have lots); 3) negatively-skewed practice (i.e., most participants have lots of practice, but some have little); and 4) bimodal practice (i.e., most participants have either little practice or lots of practice). We then investigated the effects of these different distributions of practice amounts on the correlations between the attention control measures and fluid intelligence and working memory capacity.

As shown in Figures 9a and 9b, correlations between the attention control measures and fluid intelligence and working memory capacity were similar regardless of the distribution of practice characterizing the sample. The density plots can be interpreted like a histogram and are based on 1,000 iterations for each practice distribution. The fact that the density plots for each practice distribution are nearly perfectly overlapping indicates that practice distributions did not

have a systematic effect on the overall correlation between attention control performance and fluid intelligence or working memory capacity; this can also be seen in the dashed vertical lines within the density plots depicting the mean correlation, which hardly varied as a function of the distribution of practice. In other words, the attention control tests appear robust to different participants having different amounts of practice, at least in terms of how the measures relate to fluid intelligence and working memory capacity.

Figure 9a

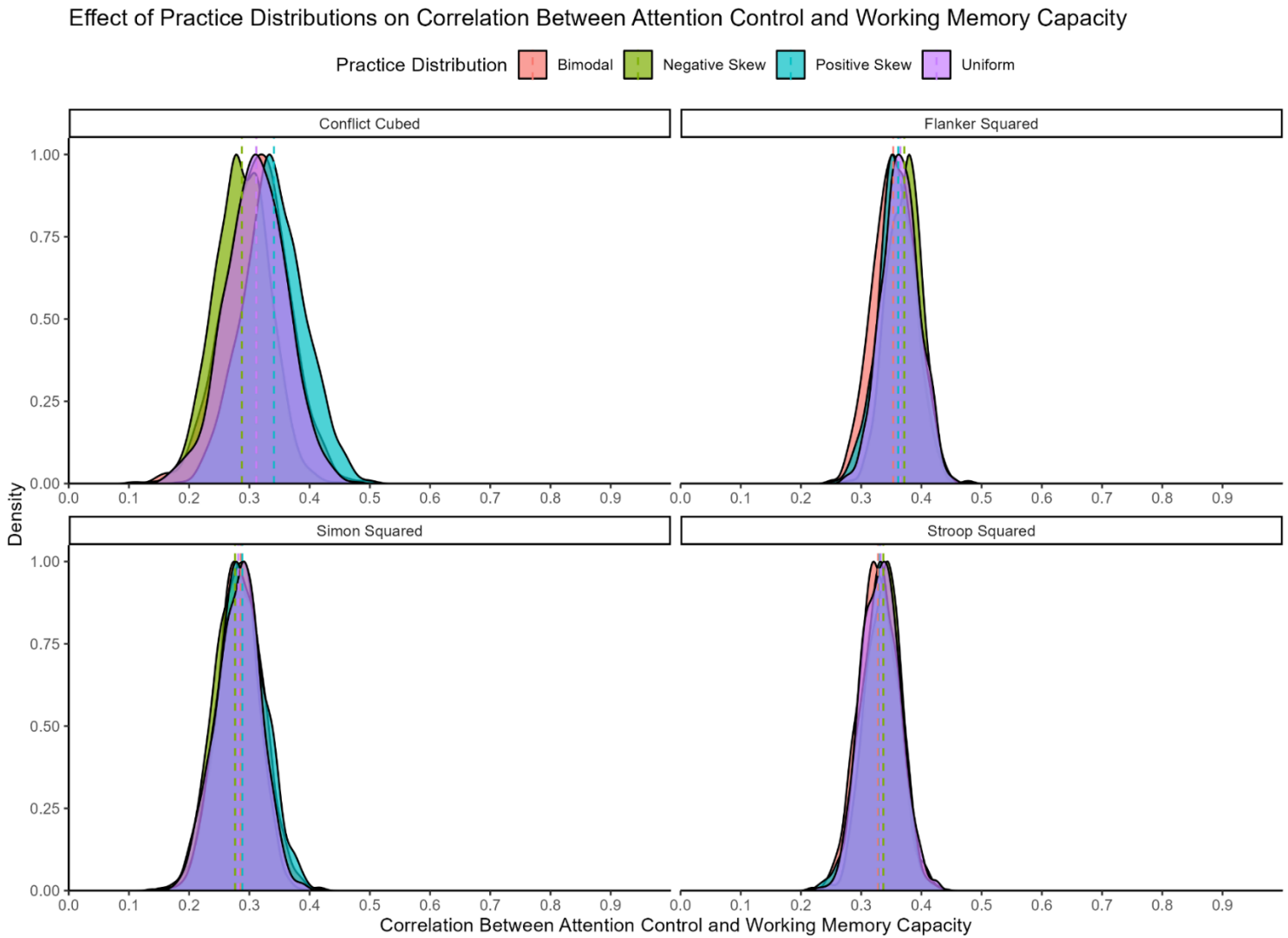
Density plots depicting effect of practice distributions on correlation between attention control and fluid intelligence



Note. The dashed vertical lines indicate the mean for each distribution.

Figure 9b

Density plots depicting effect of practice distributions on correlation between attention control and working memory



Note. The dashed vertical lines indicate the mean for each distribution.

Study 1 Discussion

In Study 1, we had a large sample of participants practice the Squared attention control tests and a new combined attention control test called Conflict Cubed to examine the effects of practice and goal reminders on the psychometric properties and predictive validity of the measures. Before and after repeated administrations, the tasks' split-half reliability estimates were .99, although this is likely a consequence of the speeded-test format. More importantly, test-retest reliability was adequate from the first attempt to the second attempt (avg. r = .71) and tended to increase when comparing performance on the second and third attempt (see Figure 5). This suggests that providing subjects with additional practice by extending the practice phase from 30s to 60-90s could improve rank-order stability. With that being said, test-retest reliability did decrease from the first attempt to later attempts; relative to performance on attempt 1, test-retest reliability fell below .7 on attempt five for Conflict Cubed, attempt 4 for Flanker Squared, attempt 3 for Stroop Squared, and attempt 2 for Simon Squared.

Participants significantly improved on every task with practice except Simon Squared, which was near ceiling performance from the outset (see Figures 6a, 6b, and 6c). The average improvement in performance from the first attempt to the second attempt across all four tasks was d = .265. Comparing the first attempt to the third attempt, the average gain score was d = .375, and these gain scores increased on subsequent attempts (see Figure 7). Thus, participants improved with practice on most tasks, which could pose consequences for relations with fluid intelligence and working memory capacity. It therefore came as a surprise that correlations between attention control and the cognitive ability factor scores representing fluid intelligence and working memory capacity did not significantly decrease with practice (see Figures 8a and 8b for the trend, see Table 3 for the tests of statistical significance).

We simulated what would happen to the correlations between attention control and fluid intelligence and working memory capacity if different participants had different amounts of practice on the attention control tests. We pseudo-randomly sampled performance attempts from participants to create new datasets characterized by different distributions of practice. Specifically, we simulated a negatively skewed practice distribution, a positively skewed practice distribution, a uniform practice distribution, and a bimodal practice distribution. Correlations with fluid intelligence and working memory capacity were robust to this manipulation (see Figures 9a and 9b). When participants had different amounts of practice, the mean correlation computed over 1,000 iterations of each practice distribution deviated by no more than $r = .04$ from the initial correlation observed when everyone was on their first attempt. In other words, participants can improve on most of the Squared tasks and Conflict Cubed with practice, but it is difficult to overcome individual differences associated with cognitive ability, at least when practice amounts are assigned at random.

Finally, the effect of goal reminders was not significant and did not interact with working memory capacity in the prediction of performance (see Table 3). We had hypothesized that providing on-screen goal reminders (e.g., “match the color of the top word to the meaning of the words below”) during testing might improve participants’ performance when first learning the task and might also reduce the role of working memory by offloading the responsibility of goal maintenance to the task environment. This was not the case. In the Introduction, we stated that if between-subjects variance in attention control tasks primarily reflects conflict resolution (as opposed to goal maintenance), then goal reminders ought to have little effect on the measures, and this is what we found.

These results are in contrast to Hood et al. (2021), who found that goal reminders reduced the Stroop-WMC correlation. However, in Hood et al.'s (2021) procedure, the task was paused intermittently so that participants could verbally rehearse the instructions before continuing. In our version, goal reminders were displayed throughout the test trials, and participants were not told to rehearse them. Thus, it is possible that goal reminders provide a facilitatory effect when offered during breaks between test trials and emphasized through verbal rehearsal, consistent with Hood et al. (2021), but have no effect or even cause distraction when presented within the test trials and without additional time for comprehension or rehearsal. In addition, Hood et al.'s (2021) version of the Stroop paradigm included mostly congruent trials, whereas the attention control tasks used here included mostly incongruent trials (see the Method for details). This could reduce the effect of goal reminders, because goal neglect is more likely to occur when trials do not directly reinforce the goal (i.e., congruent trials, Kane et al., 2003).

Study 2

In Study 2, we assessed the validity of the attention control measures and the impact of practice in a sample of U.S. Naval Flight Students: officers training to be Naval Aviators or Naval Flight Officers. This is a population of interest should the Squared tasks be used for personnel selection. We used the Naval Flight Students' academic and flight grades during initial flight training as criterion measures. Due to logistical constraints, Study 2 was limited to the Flanker Squared task administered four times. We chose Flanker Squared because previous studies conducted at the U.S. Naval Research Laboratory indicated that it shows the most promise among the three Squared tasks for aviator personnel selection (e.g., Coyne et al., 2024).

Study 2 Method

A total of 219 U.S. Naval Flight Students volunteered to complete Flanker Squared four times before beginning the NIFE program. According to the U.S. Navy, NIFE is designed to establish "…the foundations of aviation fundamentals for aspiring aviators but is also a screening tool that will test students' ability to handle stressful evolutions in a high impact environment" (America's Navy, 2024). NIFE is the initial step of Naval flight training and consists of four phases, the second and third of which are relevant to this investigation. The second phase of NIFE is a three-week-long academic phase in which students take classes and exams across five subjects (aerodynamics, engines, navigation, weather, and flight rules and regulations) and receive a percentage grade (i.e., ranging 0 – 100) based on their performance, unless they fail or attrite, in which case they receive no grade. The third phase tests students' ability to learn flight procedures and demonstrate their aeronautical adaptability. Specifically, in the third phase students are asked to "quickly memorize and prioritize information required to complete a series of flights in a Cessna 172 using Navy flight procedures…[and] attempt at least six landings and perform a set of standardized maneuvers on each of seven flights" (America's Navy, 2024). Scores for this phase of NIFE are standardized to a mean of 50 and *SD* of 10.

Naval Flight Students represent a down-selected group of individuals who qualified for NIFE based on specific criteria. These criteria include having a bachelor's degree from an accredited four-year institution and performing sufficiently well on the Aviation Selection Test Battery (ASTB), which consists of subtests assessing math skills, reading comprehension, mechanical knowledge, aviation and nautical knowledge, psychomotor ability, spatial ability, attention (via a dichotic listening test), and multitasking ability (Naval Aerospace Medical Institute, 2024). The Aviation Selection Test Battery has three composite scores which are used to predict the different phases of flight training for the student aviators and flight officers. As this

study is focused on NIFE performance, we used the Academic Qualifications Rating (AQR), a composite score which is standardized and transformed into 1-9 stanine scores for selection purposes. The AQR is the ASTB composite used to predict NIFE outcomes. The AQR stanine score is thus the appropriate comparison measure to assess the potential incremental validity of Flanker Squared in predicting training outcomes.

The version of Flanker Squared used in Study 2 was slightly different from Study 1 in that no goal reminders were presented during the test phase, and the instructions and practice were modified to improve participants' understanding. The instructions highlighted the arrows that participants should attend to in the stimulus and response options by having a blinking circle appear around them. Furthermore, during the 30s practice phase, if participants incorrectly responded, the timer stopped, the appropriate arrows in the stimulus and correct response option were highlighted, and instructional text explained the correct answer. Participants had to select the correct answer for the practice phase and timer to resume. The 90s scored portion of the test was the same as Study 1. Each Flanker Squared attempt took around three minutes, including a four-item motivation survey given after each attempt which asked participants to report on a Likert-type scale of 1-7 how fun the task was, how interesting it was, how hard they tried on it, and how important it was for them to do well on it.

Data were collected in groups of up to 15 participants at the Naval Aerospace Medical Institute located on Naval Air Station Pensacola. We aimed for a sample size of at least 200 to facilitate correlational analyses (e.g., Schönbrodt & Perugini, 2013). Participants' workstations were separated by partitions. Participants were volunteers and thus not directly compensated but were informed that their participation could help improve future aviator selection practices. The protocol was approved by the U.S. Naval Research Laboratory's Institutional Review Board.

Transparency and Openness

We describe our data exclusions, all manipulations, and all measures in the study. Data for Study 2 are kept and protected by the U.S. Navy (specifically, the Naval Aerospace Medical Institute and the Naval Research Laboratory) and can only be shared if the requester can demonstrate to all relevant parties that required data security protocols will be adhered to. Data were analyzed using R, version 4.0.0 (R Core Team, 2020).

Data Preparation

Of the 219 participants who provided informed consent, six were removed from the dataset for not completing the NIFE academic phase and thus not having NIFE academic grade or flight score, resulting in a final *N* of 213. Of these 213, three had a Flanker Squared score at or below 5 points, and in all cases, it was their first attempt. These three attempt 1 scores were labeled as missing (as were all of their responses in the trial-level data), and scores on all other attempts were retained. Average RT for each participant on each Flanker Squared attempt was under 8 seconds; no scores were excluded due to slow RTs. Of the 213 participants in the final dataset, 849 out of 852 (99.6%) of their total scores were retained. However, it was not possible to link AQR and NIFE flight scores to all participants, so analyses involving AQR have $n = 208$ and analyses involving NIFE flight scores have $n = 180$. Missing AQR scores were due to missing and/or conflicting information in the demographic records and the large number of missing flight scores was because some students are allowed to bypass the flight portion of NIFE if they already had similar training, such as Naval Academy Cadets who previously participated in their Powered Flight Program.

Study 2 Results

Table 5 provides descriptive statistics for each Flanker Squared attempt, AQR stanine scores, NIFE academic grades, and NIFE flight grades. Table 6 reports gain scores and internal consistency reliability on each attempt. On average, Flanker Squared scores increased from 39.44 on the first attempt to 46.94 by the fourth. In terms of Cohen's *d,* improvements were d = 0.25 from attempt one to two, d = 0.44 from one to three, and d = 0.73 from one to four.

Table 7 shows self-reported motivation responses after each Flanker Squared attempt and averaged across all four attempts, and Table 8 shows the Pearson correlations between the aggregated responses for each of the four motivation questions (i.e., fun, interesting, tried, and important). We combined responses on the questions pertaining to participant effort and the importance of performing well to create a motivation score. Mean values on this motivation score were both consistent and high across Flanker Squared attempts (M = 10.42 – 10.75 across attempts; note that 14 was the highest possible score). Attempt-specific motivation scores correlated with each other between r = .79 - .92 (all ps < .05), and correlated sporadically with performance on their respective attempts: Flanker 1 motivation did not correlate with Flanker 1 performance (r = .05, p = .453); Flanker 2 motivation correlated with Flanker 2 performance (r = .17, p = .012); Flanker 3 motivation correlated with Flanker 3 performance (r = .17, p = .016); and Flanker 4 motivation did not correlate with Flanker 4 performance (r = .11, p = .103).

Table 9 shows the bivariate correlations between all Flanker attempts, criterion measures, and AQR. Each Flanker Squared attempt correlated significantly with NIFE academic grades: Flanker 1 (r = .32, p < .001), Flanker 2 (r = .25, p < .001), Flanker 3 (r = .20, p = .004), and Flanker 4 (r = .26, p < .001). The strongest correlation was observed on the first attempt and the weakest on the third. Using Steiger's (1980) Z test for the difference between dependent

correlations, the correlation between the first Flanker Squared attempt and NIFE academic grade was statistically larger than the correlation between the third attempt and NIFE academic grade (Z_h = 2.09; p = .036). None of the other correlations involving Flanker Squared scores and NIFE academic grade were statistically different from one another (all ps > .10). For NIFE flight scores, the correlation between Flanker Squared and flight performance was statistically significant on the first (r = .15, p = .044) and fourth attempt (r = .18, p = .014). Correlations between Flanker Squared and AQR stanine scores ranged from r = .22 to r = .25.

Table 5

Descriptive Statistics for Study 2 Performance Variables

Variable	*n*	*M*	*SD*	Min - Max	Skew	Kurtosis
Flanker Squared 1	210	39.44	10.30	8 - 74	-0.08	0.70
Flanker Squared 2	213	42.05	11.00	14 - 74	0.08	0.25
Flanker Squared 3	213	43.99	11.31	7 - 72	-0.15	-0.05
Flanker Squared 4	213	46.94	11.05	19 - 79	-0.53	-0.26
AQR Stanine	207	6.74	1.29	4 - 9	0.01	-0.81
NIFE Academic	213	94.32	3.70	81.20 - 100	-0.96	0.73
NIFE Flight	180	49.77	9.24	34.18 – 83.77	1.10	1.45

Note. AQR = Academic Qualification Rating Stanine; NIFE = Naval Introductory Flight Evaluation program. NIFE Academic grades were a percentage that could range from 80 – 100 as below 80 is considered a failing grade. NIFE Flight scores are standardized to a mean of 50 and *SD* of 10.

Table 6

Gains and Reliability of Flanker Squared in Study 2

Variable	Gains (*d*)	Estimated Reliability
Flanker Squared 1	0.00	.978
Flanker Squared 2	0.25	.980
Flanker Squared 3	0.44	.981
Flanker Squared 4	0.73	.981

Note. Gains were indexed as Cohen's *d* and were computed using the first attempt *SD* as the standardizer for each attempt. Internal consistency reliability was estimated using even vs. odd numbered trials and adjusted using the Spearman-Brown prophecy formula.

Table 7

Descriptive Statistics for Study 2 Self-Report Questions

Variable	*n*	*M*	*SD*	Min – Max	Skew	Kurtosis
Fun 1	210	4.68	1.53	1 – 7	-0.38	-0.59
Fun 2	213	4.61	1.64	1 – 7	-0.33	-0.59
Fun 3	213	4.59	1.70	1 – 7	-0.29	-0.67
Fun 4	213	4.66	1.71	1 – 7	-0.34	-0.79
Fun Average	213	4.63	1.54	1 – 7	-0.30	-0.58
Interesting 1	207	4.23	1.67	1 – 7	-0.40	-0.71
Interesting 2	208	4.36	1.62	1 – 7	-0.16	-1.03
Interesting 3	209	4.35	1.67	1 – 7	-0.17	-0.84
Interesting 4	209	4.42	1.67	1 – 7	-0.29	-0.82
Interesting Average	209	4.34	1.57	1 – 7	-0.15	-0.76
Tried 1	205	5.46	1.31	1 – 7	-0.91	0.60
Tried 2	208	5.28	1.42	1 – 7	-0.80	0.13
Tried 3	208	5.27	1.51	1 – 7	-0.87	0.31
Tried 4	207	5.32	1.50	1 – 7	-0.83	0.17
Tried Average	208	5.33	1.32	1 – 7	-0.79	0.29
Important 1	204	5.30	1.53	1 – 7	-0.93	0.29
Important 2	207	5.22	1.54	1 – 7	-0.83	0.47
Important 3	205	5.17	1.60	1 – 7	-0.81	0.01
Important 4	206	5.28	1.56	1 – 7	-0.87	0.25
Important Average	207	5.24	1.49	1 – 7	-0.83	0.23
Motivation 1	204	10.75	2.54	2 – 14	-0.73	0.13

Motivation 2	207	10.50	2.75	2 – 14	-0.68	-0.12
Motivation 3	205	10.42	2.95	2 – 14	-0.79	0.19
Motivation 4	206	10.61	2.90	2 – 14	-0.78	0.25
Motivation Average	207	10.56	2.63	2 – 14	-0.70	0.08

Note. The digit after each measure refers to the Flanker Squared attempt to which it pertains. "Average" is the average for each participant across all attempts. "Motivation" is the sum of tried plus important. Averages were not calculated for participants who were missing more than one data point for the respective question (e.g., if a participant did not answer the fun question after attempts 1 and 2, then their "Fun Average" was not calculated).

Table 8

Correlations among Study 2 Self-Report Questions

Variable	Fun	Interesting	Tried
Fun	---		
Interesting	**.88**	---	
Tried	**.63**	**.60**	---
Important	**.65**	**.61**	**.76**

Note. All correlations were significant at $p < .05$. Each self-report question was administered after each four Flanker Squared attempts; this table shows the correlations among the average responses across all attempts.

Table 9

Correlations among Study 2 Variables

Variable	Flanker 1	Flanker 2	Flanker 3	Flanker 4	AQR	NIFE Academic	NIFE Flight
Flanker 1	---						
Flanker 2	**.72**	---					
Flanker 3	**.69**	**.76**	---				
Flanker 4	**.70**	**.77**	**.80**	---			
AQR Stanine	**.25**	**.22**	**.22**	**.22**	---		
NIFE Academic	**.32**	**.25**	**.20**	**.26**	**.43**	---	
NIFE Flight	**.15**	.14	.11	**.18**	**.29**	**.33**	
Motivation	.01	.13	.12	.10	**.21**	.03	.10

Note. Correlation values in bold were significant at $p < .05$. AQR = Academic Qualification Rating Stanine; NIFE = Naval Introductory Flight Evaluation program; Motivation = sum of "tried" plus "important" self-report questions averaged across all four attempts.

Number of Trials

Considering the 213 participants and 849 task scores that entered into the analyses, participants completed an average of 187.81 total trials (*SD* = 36.79; min = 77; max = 287) across all attempts. The mean number of trials participants completed on Flanker Squared attempt 1 was 43.75 (*SD* = 8.75; min = 20; max = 74); for attempt 2, the number of trials completed was *M* = 46.05 (*SD* = 9.64; min = 23; max = 77); for attempt 3, the number of trials completed was *M* = 48.01 (*SD* = 10.02; min = 26; max = 74); and for attempt 4, the number of trials completed was *M* = 50.61 (*SD* = 10.52; min = 25; max = 79).

Practice Distribution Simulations

Mirroring Study 1, we simulated four different practice distributions (i.e., positively skewed, negatively skewed, uniform, and bimodal) by sampling a specific Flanker Squared

attempt for each participant in the dataset, computing the correlation with NIFE academic grade or NIFE flight scores, and repeating this process 1,000 times for each distribution.

NIFE Academic Grade. Across the four practice distributions, the average correlation between Flanker Squared performance and NIFE academic grade ranged from r = .23 to .26, with an average standard deviation of .03 (Table 10). For comparison, refer to Table 9 in which the strongest correlation between Flanker Squared and NIFE grades was r = .31 when everyone was on their first attempt, and the weakest correlation was r = .20 when everyone was on their third attempt. Thus, the predictive validity of the Flanker Squared was not considerably reduced when participants had different amounts of practice, and the results were similar regardless of which practice distribution characterized the sample.

Table 10

Effect of Practice Distributions on Correlation between Flanker Squared and NIFE Academic Grade

Practice Distribution	Average Correlation	Min - Max	*SD*
Uniform	.24	.13 - .35	.03
Positive Skew	.25	.14 - .35	.03
Negative Skew	.23	.14 - .31	.03
Bimodal	.26	.16 - .36	.03

Note. The attempt-sampling method was performed 1,000 times for each practice distribution. Uniform = each attempt is equally likely to be chosen; Positive Skew = early attempts are more likely to be chosen; Negative Skew = later attempts are more likely to be chosen; Bimodal = the first or last attempts are more likely to be chosen.

NIFE Flight Scores. We also performed the same practice distribution simulations for NIFE flight scores (Table 11). In this set of analyses, the average correlation between Flanker Squared performance and NIFE flight scores ranged from r = .14 to .15, each with a standard deviation of .04. For comparison, refer to Table 9 in which the strongest correlation between Flanker Squared and NIFE grades was r = .18 (fourth attempt) and the weakest correlation was r = .11 (third attempt). Thus, the predictive validity of the Flanker Squared was not considerably reduced when participants had different amounts of practice, and the results were similar regardless of which practice distribution characterized the sample.

Table 11

Effect of Practice Distributions on Correlation between Flanker Squared and NIFE Flight Scores

Practice Distribution	Average Correlation	Min - Max	*SD*
Uniform	.14	.01 - .26	.04
Positive Skew	.14	.01 - .25	.04
Negative Skew	.15	.02 - .26	.04
Bimodal	.15	.01 - .28	.04

Note. The attempt-sampling method was performed 1,000 times for each practice distribution. Uniform = each attempt is equally likely to be chosen; Positive Skew = early attempts are more likely to be chosen; Negative Skew = later attempts are more likely to be chosen; Bimodal = the first or last attempts are more likely to be chosen.

Commonality Analyses for Predictive Validity of AQR and Flanker Squared

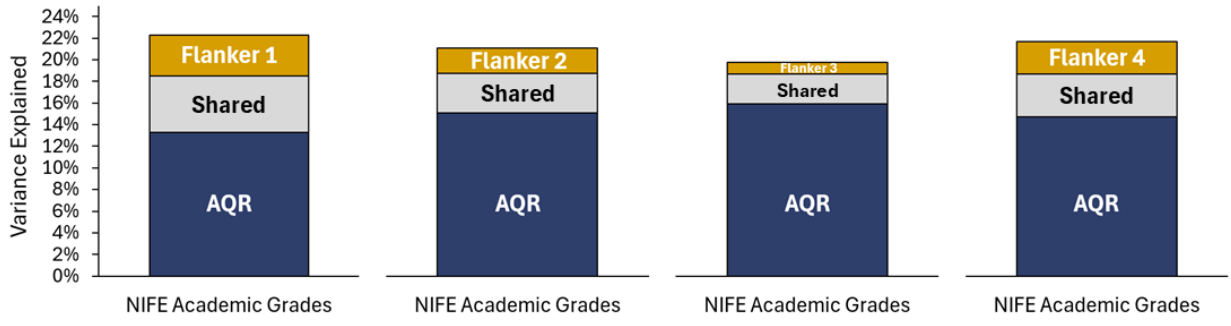
Our final set of analyses concerned the extent to which each Flanker Squared attempt provided incremental validity in the prediction of training outcomes (i.e., NIFE academic grades and flight scores) above and beyond AQR scores. We performed two separate regression

analyses to obtain commonality coefficients (Nimon & Oswald, 2013), partitioning variance in NIFE grades into three parts: (1) variance shared by both Flanker Squared and AQR, (2) variance uniquely explained by AQR, and (3) variance uniquely explained by Flanker Squared. The third component represents the incremental validity of Flanker Squared in predicting NIFE academic grades or flight scores above and beyond AQR.

NIFE Academic Grade. As shown in Figure 10, AQR and Flanker Squared together significantly predicted NIFE academic grades, ranging from predicting 19.5 – 21.8% of the total variance depending on the Flanker Squared attempt. Breaking down the relative contributions of AQR and Flanker Squared, they both accounted for unique variance in NIFE academic grades. For Flanker Squared, the unique proportion of variance explained was 4.1% (attempt one), 2.3% (attempt two), 1.1% (attempt three), and 3.0% (attempt four), all of which were statistically significant except for attempt three. AQR explained between 12.8 - 15.7% unique variance, depending on the Flanker Squared attempt. The overlap (i.e., shared variance among both predictors and NIFE academic grade) ranged from 2.7 - 5.2%.

Figure 10

Commonality analysis with Flanker Squared and AQR predicting NIFE academic grades



Note. The height of each bar indicates the proportion of variance in NIFE grades that was explained by each predictor. The gray shaded region in the middle represents variance that was jointly explained (i.e., shared) by Flanker Squared and AQR.

NIFE Flight Scores. For NIFE flight scores, the combined variance explained by AQR and Flanker Squared was substantially less than it was for NIFE academic grade, ranging from 8.7 – 10.1% depending on Flanker Squared attempt. Further, while AQR accounted for statistically significant unique variance (6.1 - 7.3%), Flanker Squared did not (0.3 - 1.6%). There was also less variance jointly explained by both predictors (1.1 - 2.1%, depending on the Flanker Squared attempt).

Study 2 Discussion

Compared to Study 1, average scores on Flanker Squared were substantially higher in Study 2 (e.g., 28 points vs. 39 points on attempt one). This was to be expected given the specialized and highly motivated Navy population used in this study. Importantly, however, standard deviations did not decrease with practice. The correlation between the *SD* and attempt number trended in the opposite direction ($r = .77$, $p = .23$, $n = 4$), although it was not statistically significant. With respect to performance improvements due to practice, gains were around 40% larger in Study 2 than Study 1, with a total improvement of $d = 0.73$ from the first to fourth attempt in Study 2 versus $d = 0.52$ in Study 1. Several factors may have contributed to this, including in-lab group testing, Naval Flight Students being a high-ability population especially motivated to perform well, and the improvements that were made to the task's practice phase.

At the bivariate level, Flanker Squared explained roughly 4% to 10% of the variance in NIFE academic grades, depending on the attempt. Regression-based commonality analyses

further revealed that, depending on the attempt, 2.7% to 5.2% of the variance explained in NIFE academic grades was jointly explained by Flanker Squared and AQR stanine. Flanker Squared explained between 1.1% to 4.1% unique variance in NIFE academic grades above and beyond AQR. A 90-second test explaining 10% of the variance in a three-week academic training program is impressive, especially given the range restriction of this sample and the low attrition rate observed in NIFE (3.2% in this dataset, and all but one attrite occurring before NIFE academic course grades were assigned; see Sibley [2024] for other subtest correlations with training outcomes). For comparison, AQR stanine scores are derived from a comprehensive battery administered over three hours and explained around 18% of the variance in NIFE academic grades at the bivariate level, of which between 12.8% to 15.7% was unique after accounting for Flanker Squared, depending on the attempt. Interestingly, Flanker Squared performance only correlated statistically significantly with NIFE flight scores on attempts 1 and 4 (r = .15 and .18, respectively). We note that the sample size for NIFE flight scores was lower compared to other variables in the dataset, which reduced the statistical power associated with these effects.

One of the primary goals of Study 2 was to assess the impact of Naval Flight Students having different amounts of practice on the predictive validity of Flanker Squared. This analysis simulates a real-world selection concern: in the modern internet era in which test information is exceedingly challenging to contain, candidates will almost certainly take selection tests with different levels of familiarity, practice, and knowledge of effective strategies. It is therefore desirable for a test to be as robust as possible to such effects while still meaningfully contributing to the prediction of relevant outcomes. From one perspective, the simulations of Tables 10 and 11 are encouraging, as regardless of the practice distribution we sampled, Flanker Squared scores

and NIFE outcomes were correlated around the same level as when everyone had the same amount of practice (i.e., r = .23 - .26 for Flanker Squared performance and NIFE academic grades; r = .14 - .15 for Flanker Squared performance and NIFE flight scores). But across practice distributions, while the correlations between Flanker Squared and NIFE academic grade followed a normal distribution with most iterations falling within ± .03 of the mean, it is noteworthy that the range for each of the 1,000 iterations was fairly large, with a minimum of r = .13 and a maximum of r = .37 depending on the random sample used to simulate each practice distribution. In other words, Flanker Squared could explain anywhere between around 2% to 14% of the variance in NIFE academic grades depending on the simulated distribution of practice among examinees. However, the vast majority of random samples yielded correlations similar to the average value (we refer the reader to Figures 9a and 9b for density plots showing the distribution of correlations that emerged in Study 1). Should these tasks be used for personnel selection, candidates having different levels of experience could introduce a meaningful amount of noise and uncertainty into the selection process. This "noise" is not just theoretical or hypothetical; it is a current threat to validity that affects many selection tests being used by the U.S. military today. Therefore, we encourage researchers to take a similar approach to the one used here to explore and better understand the effects of practice and different people having different amounts of practice on military selection tests.

The other goal of Study 2 was to assess the incremental validity of Flanker Squared in predicting NIFE outcomes above and beyond AQR, the standardized score from the Aviation Selection Test Battery used to screen and select applicants into Naval Flight School. Flanker Squared performed well in regard to incremental validity in predicting NIFE academic grades, adding as much as 4.1% variance above and beyond AQR. However, this was not the case for the

prediction of NIFE flight scores, in which incremental validity of Flanker Squared over AQR was not statistically significant for any of the four attempts. It is likely that had performance been measured at the latent level with multiple indicator measures, instead of using single indicator observed scores, the incremental variance explained by performance on the attention control tests might be greater.

General Discussion

Across two studies (combined $N > 700$), we assessed how the psychometric properties of attention control measures were affected by practice, and by different participants having different amounts of practice. Overall, we found that scores on every task except Simon Squared improved significantly with practice, but practice had little effect on the measures' relationships with other cognitive abilities and U.S. Naval Aviator training performance. Novel to the present studies, we simulated what would happen if different test takers had different amounts of experience and found that criterion-related validity was robust to four different distributions of practice. The overall conclusion to emerge from this work is that people can improve with practice on attention control tests, but that individual differences in cognitive ability are difficult to overcome.

We found no evidence for range restriction as a result of subjects gaining experience on the tasks. In Study 1, standard deviations significantly increased from the first attempt to the last attempt on all four tasks, consistent with Xu et al. (2018). In Study 2, the relationship between the standard deviation and attempt number was positive but not statistically significant ($r = .77$, $p = .23$). Thus, our results are in contrast with theories of skill acquisition that predict decreasing ability-related differences with experience, such as the circumvention of limits hypothesis. For a

large-scale test of this hypothesis which also showed little convergence of ability groups over time, see Hambrick et al. (2024).

In the real world, practice distributions are typically positively skewed (Hambrick et al., 2014; Gobet & Campitelli, 2007). Our study provides a general framework and analytical approach to assess the consequences of different distributions of practice on ability measures. Somewhat surprisingly, we found that whether the amount of practice within a sample followed a positively skewed, negatively skewed, uniform, or bimodal distribution, the results were largely similar. However, it is possible that with even more practice, a different pattern of results could emerge. In Study 1, participants completed an average of 1249 trials on Conflict Cubed and on average of 406-561 trials on each of the Squared tasks, whereas in Study 2, they completed an average of 188 trials on Flanker Squared. Compared to studies like Paap et al. (2014), in which a participant completed over 20,000 trials and was still demonstrating a reduction in the conflict effect with practice, it is clear that the tail of the practice distribution could be much longer. Therefore, one explanation for the relatively consistent relations we observed between the practiced tasks and outcome measures is that more practice is required on the complex tasks we administered here to develop the level of automaticity required to observe changing ability-performance relations. Importantly, because we did not observe a significant practice × cognitive ability interaction, individual differences in performance were largely preserved across attempts. As a result, reshuffling which attempt represented each participant's score was less likely to influence their rank order, and in turn the validity of the measures. Thus, a future direction for this work is to give participants as much practice as possible to map the full learning curve through its asymptote.

Although we found relatively stable relations with the attention control tasks as a

function of practice, this would probably not be the case for many cognitive ability tests (see, e.g., Sibley et al., 2023). It would be worthwhile to extend this approach to other tests that demonstrate reduced correlations with outcome measures following extensive practice or coaching; under this circumstance, we would expect different distributions of practice to have different effects on criterion-related validity. Another extension of this work is to simulate the consequences of practice on the measures' validity when cut scores are used to make selection decisions. Our results focused on overall validity statistics (e.g., correlations with Aviator training performance), but practice could allow a participant to eke out a score above a selection cut point, which would affect selection rates. Under this circumstance, smaller improvements in performance could result in larger differences in outcomes.

We plan to address a major limitation of our approach in future work by simulating what would happen if different characteristics of the participants were associated with their likelihood to practice, especially if those characteristics were related to ability. For example, if initial performance predicts how much a participant decides to practice, then it could lead to a "rich-get-richer" effect if the interaction term is positive or a compensatory effect for lower-performers if the interaction term is negative. As another example, if participants with greater openness to experience are more likely to practice, and openness correlates with cognitive ability (e.g., DeYoung et al., 2014; Stanek & Ones, 2023), then we would expect to see a widening of the ability gap driven by participants with greater openness. In short, our simulations did not account for individual differences that are likely to affect the propensity to practice, but these differences exist and could be modeled in future studies.

One of the critical questions that remains to be addressed concerns the meaning of the measure of performance provided by scores on attention control tests (or other cognitive ability

tests) following extensive practice. Many tests of attention control, such as the Stroop task, require inhibiting a prepotent or habitual response (e.g., inhibiting reading the word "RED" aloud, as opposed to stating the color it appears in). However, with increasing exposure to the test, it is likely that participants develop test-specific strategies that help them reduce the interference effect and respond with greater accuracy and speed. Theoretically, with enough practice, responses to high-conflict (i.e., incongruent) attention control trials could be governed by automatic processes (e.g., Fitts & Posner, 1967) and would no longer tap the ability to inhibit prepotent responses to the same degree as initial performance attempts. Despite this possibility, the attention control measures studied here remained predictive even after extensive practice, but that does not mean that the measures continue to tap the same attention control on attempt 10 (or, hypothetically, attempt 1000) as on attempt 1. It is possible that individual differences in task-specific strategies become increasingly important to performance, but the discovery and implementation of these strategies is related to cognitive ability. If so, then the measure could continue to predict outcomes despite tapping a fundamentally different ability. One correlational approach to testing this would be to examine relations with other established measures of attention control prior to and following extensive practice on the tasks; a reduction in the correlation would suggest that non-attention-control-related processes are playing more of a role following extensive practice.

Supportive evidence for this possibility is provided by a training study involving two-alternative forced-choice processing speed measures conducted by Reinhartz et al. (2023). In their study, Reinhartz et al. (2023) used drift diffusion modeling to disentangle *drift rate* (i.e., the rate of evidence accumulation towards a decision boundary), *boundary separation* (i.e., the degree of response caution), and *non-decision time* (i.e., motoric and other processes), and

modeled how these parameters changed as a function of practice. They found that drift rate increased, suggesting that participants were quicker at accumulating evidence about the trial following extensive practice. This could reflect the development of task-specific strategies or a honing of the cognitive system improving the processing of trial-level information, allowing participants to more rapidly extract evidence for the correct response option. In a sense, this suggests that trials became easier for participants as they gained more experience on the task. They also found that boundary separation decreased, suggesting a decrease in response caution with practice, and that non-decision time decreased, suggesting that motor efficiency and other processes may have increased. Thus, if diffusion model parameters are taken as diagnostic, Reinhartz et al.'s (2023) study suggests that the efficiency of task-specific cognitive processing improves with practice, though what drives this particular improvement is unclear.

Given that the rank-order stability of participants appeared to improve after attempt one (see Figures 5 and 7), as indicated by increased test-retest reliability and reduced gain scores when treating attempt two as the baseline measure of performance, we suggest that researchers interested in using the Squared attention control tests give participants more effective practice with the tasks prior to entering the test phase. Initially, we provided participants with only 30 seconds of practice. In Study 2, the practice phase was altered so that participants could only proceed by responding correctly to the practice trials. In general, we think giving participants enough instruction so that they fully understand the task before entering the test phase would be worthwhile, as floor effects from participants misunderstanding the task on the first attempt could be reduced and retest reliability could be improved. This may entail increasing the practice phase from 30 seconds to 60 seconds, or providing error correction during practice akin to the version used in Study 2.

With respect to criterion-related validity, we found that attention control explained unique variance in NIFE academic grades in U.S. Naval Aviators in training above and beyond the AQR stanine score, the Academic Qualification Rating used to help identify future U.S. Naval Aviators for training. Although incremental validities were modest, explaining 1.1% to 4.1% of the variance beyond AQR stanine scores, this might be seen as fairly impressive considering the test took 3 minutes or less to administer, predicted performance in a 3-week long training program, did so above and beyond an extensive 3-hour assessment that covers a broad range of topics designed to predict aviation training success, and was administered to a down-selected restricted range sample of high-ability personnel.

Conclusion

To build robust assessments of cognitive ability for high-stakes tests, developers must be prepared to contend with individual differences in practice and their effects on reliability and validity. We assessed the psychometric properties and validity of three-minute "Squared" attention control tests before and after practice in a university sample and a sample of U.S. Aviators in training. The measures demonstrated strong psychometric properties and criterion-related validity following extensive practice. We also developed a novel approach to simulate different distributions of practice within a sample and examine their effects on criterion-related validity. Because practice effects and coaching should be expected on high-stakes tests, developers should thoroughly explore how more extensive practice—and variation in practice across participants—affects their assessments. We have provided a general approach to do so, using attention control measures as our test case. Our hope is that this approach will be adopted widely to better understand and ultimately safeguard high-stakes assessments against the effects of practice.

Constraints on Generality

This work focused on practice effects in samples of university participants and U.S. Naval Flight Students. As such, it is unclear whether the same conclusions would hold in distinctly different samples of participants (e.g., children or the elderly). Furthermore, it remains an open question how the results might differ if considerably more practice were administered to participants than was administered here.

Public Significance Statement

When job applicants take high-stakes cognitive tests for selection or classification, some candidates may have practiced the tests extensively while others have not. Therefore, we examined whether differences in practice would undermine the reliability and validity of three-minute "Squared" attention control tests. Across two studies including university students and U.S. Naval Flight Students, we found that while people improved with practice, the tests remained valid predictors of training performance even when test-takers had vastly different amounts of practice. These findings suggest that brief attention control tests can provide robust assessments of cognitive ability and that their validity is relatively resistant to practice effects.

References

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General*, *117*(3), 288-318.

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659-701.

America's Navy (2024). NIFE Lays Foundation for Naval Aviation Training Retrieved November 18, 2024 from: https://www.navy.mil/Press-Office/News-Stories/Article/2944668/nife-lays-foundation-for-naval-aviation-training/

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556-559.

Bates TC, Neale MC, Maes HH (2019). "umx: A library for Structural Equation and Twin Modelling in R." *Twin Research and Human Genetics*, **22**, 27-41. doi:10.1017/thg.2019.2.

Boring, E. G. (1961). Intelligence as the Tests Test It. In J. J. Jenkins & D. G. Paterson (Eds.), *Studies in individual differences: The search for intelligence* (pp. 210–214). Appleton-Century-Crofts. https://doi.org/10.1037/11491-017

Burgoyne, A. P., & Engle, R. W. (2020). Attention control: A cornerstone of higher-order cognition. *Current Directions in Psychological Science*, *29*(6), 624-630.

Burgoyne, A. P., Mashburn, C. A., Tsukahara, J. S., & Engle, R. W. (2022). Attention control and process overlap theory: Searching for cognitive processes underpinning the positive manifold. *Intelligence*, *91*, 101629.

Burgoyne, A. P., Mashburn, C. A., Tsukahara, J. S., Pak, R., Coyne, J. T., Foroughi, C., Sibley, C., Drollinger, S. M., & Engle, R. W (2024). Attention control measures improve the

prediction of performance in Navy trainees. *International Journal of Selection and Assessment,* 33: e12510. https://doi.org/10.1111/ijsa.12510

Burgoyne, A. P., Tsukahara, J. S., Mashburn, C. A., Pak, R., & Engle, R. W. (2023). Nature and measurement of attention control. *Journal of Experimental Psychology: General, 152*(8), 2369–2402. https://doi.org/10.1037/xge0001408

Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist*, *26*(4), 543-570.

Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic bulletin & review*, *12*, 769-786.

Coyne, J. T., Brown, N. L., Foroughi, C. K., Sibley, C., Sexauer, E., & Rovira, E. (2020, December). The use of non-spatial strategies in the Direction Orientation Task. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 64, No. 1, pp. 802-806). Sage CA: Los Angeles, CA: SAGE Publications.

Coyne, J. T., Draheim, C., Sibley, C., Foroughi, C., Strong, K., NeSmith, R., Burgoyne, A. P., & Engle, R. W. (2024, April). *Evaluation of short attention tests for selecting Navy air traffic controllers*. Poster presented at the 39th Society for Industrial and Organizational Psychology annual conference, Chicago, IL, United States.

Coyne, J. T., Draheim, C., Sibley, C., Armendariz, N., Melick, S., Foroughi, C., Burgoyne, A. P., & Engle, R. W. (2024). Evaluation of New Measures of Spatial Ability and Attention Control for Selection of Naval Flight Students. *Human Factors in Design, Engineering, and Computing, Vol. 159*, 2024, 2006–2016. https://doi.org/10.54941/ahfe1005768

DeYoung, C. G., Quilty, L. C., Peterson, J. B., & Gray, J. R. (2014). Openness to experience, intellect, and cognitive ability. *Journal of Personality Assessment*, *96*(1), 46-52.

Draheim, C., Harrison, T. L., Embretson, S. E., & Engle, R. W. (2018). What item response theory can tell us about the complex span tasks. *Psychological Assessment*, *30*(1), 116.

Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological Bulletin*, *145*(5), 508.

Draheim, C., Tsukahara, J. S., Martin, J. D., Mashburn, C. A., & Engle, R. W. (2021). A toolbox approach to improving the measurement of attention control. *Journal of Experimental Psychology: General*, *150*(2), 242.

Duncan, J., Emslie, H., Williams, P., Johnson, R., & Freer, C. (1996). Intelligence and the frontal lobe: The organization of goal-directed behavior. *Cognitive Psychology*, *30*(3), 257-303.

Ekstrom R. B., French J. W., Harman H. H., & Dermen D. (1976). *Manual for kit of factor-referenced cognitive tests: 1976*. Educational Testing Service.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of experimental psychology: General*, *128*(3), 309.

Fitts, P.M., & Posner, M.I. (1967). *Human performance.* Brooks/Cole.

Fleishman, E. A. (1972). On the relation between abilities, learning, and human performance. *American Psychologist*, *27*(11), 1017.

Fleishman, E. A., & Hempel Jr, W. E. (1954). Changes in factor structure of a complex psychomotor test as a function of practice. *Psychometrika*, *19*(3), 239-252. https://doi.org/10.1007/BF02289188

Fleishman, E. A., & Rich, S. (1963). Role of kinesthetic and spatial-visual abilities in perceptual-motor learning. *Journal of Experimental Psychology*, *66*(1), 6. https://psycnet.apa.org/doi/10.1037/h0046677

Gobet, F., & Campitelli, G. (2007). The role of domain-specific practice, handedness, and starting age in chess. *Developmental psychology*, *43*(1), 159.

Hambrick, D. Z., Burgoyne, A. P., & Oswald, F. L. (2024). The validity of general cognitive ability predicting job-specific performance is stable across different levels of job experience. *Journal of Applied Psychology, 109*(3), 437–455. https://doi.org/10.1037/apl0001150

Hambrick, D. Z., Oswald, F. L., Altmann, E. M., Meinz, E. J., Gobet, F., & Campitelli, G. (2014). Deliberate practice: Is that all it takes to become an expert?. *Intelligence*, *45*, 34-45.

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: a meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, *92*(2), 373.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior research methods*, *50*, 1166-1186.

Hood, A. V., Charbonneau, B., & Hutchison, K. A. (2022). Once established, goal reminders provide long-lasting and cumulative benefits for lower working memory capacity individuals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(12), 1738.

Hood, A. V., & Hutchison, K. A. (2021). Providing goal reminders eliminates the relationship between working memory capacity and Stroop errors. *Attention, Perception, & Psychophysics*, *83*, 85-96.

Kane, M. J., & Engle, R. W. (2003). Working-memory capacity and the control of attention: the contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of experimental psychology: General*, *132*(1), 47.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: a latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of experimental psychology: General*, *133*(2), 189.

Lurie, K. (2010). *Cracking the GRE*. Princeton Review.

Mashburn, C. A., Burgoyne, A. P., & Engle, R. W. (2023). Working memory, intelligence, and life success. *Memory in Science for Society: There Is Nothing As Practical As a Good Theory*, 149.

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current directions in psychological science*, *21*(1), 8-14.

Navy Aerospace Medical Institute, (2024). ASTB-E Frequently Asked Questions. Retrieved November 18, 2024 from: https://www.med.navy.mil/Navy-Medicine-Operational-Training-Command/Naval-Aerospace-Medical-Institute/ASTB-FAQ/

Nimon, K. F., & Oswald, F. L. (2013). Understanding the results of multiple linear regression: Beyond standardized regression coefficients. *Organizational Research Methods*, *16*(4), 650-674.

Paap, K., Wagner, S., Johnson, H., Bockelman, M., Cushing, D., & Sawi, O. (2014). 20,000 Flanker Trials: Are the Effects Reliable, Robust, and Stable? [Poster abstract]. 2014 American Psychological Society Annual Meeting.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raven, J. C., & Court, J. H. (1998). *Raven's progressive matrices and vocabulary scales* (Vol. 759). Oxford: Oxford Psychologists Press.

Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012). Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment*, *28*(3), 164–171. https://doi.org/10.1027/1015-5759/a000123

Reinhartz, A., Strobach, T., Jacobsen, T., & von Bastian, C. C. (2023). Mechanisms of Training-Related Change in Processing Speed: A Drift-Diffusion Model Approach. *Journal of Cognition*, *6*(1).

Rouder, J. N., Kumar, A., & Haaf, J. M. (2023). Why many studies of individual differences with inhibition tasks may not localize correlations. *Psychonomic Bulletin & Review*, *30*(6), 2049-2066.

Schmidt, F. L., Hunter, J. E., Outerbridge, A. N., & Goff, S. (1988). Joint relation of experience and ability with job performance: Test of three hypotheses. *Journal of Applied Psychology, 73*(1), 46–57. https://doi.org/ 10.1037/0021-9010.73.1.46

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*(1), 1. https://psycnet.apa.org/doi/10.1037/0033-295X.84.2.127

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize?. *Journal of Research in Personality, 47*(5), 609-612. https://doi.org/10.1016/j.jrp.2013.05.009

Sibley, C. (2024). *High-stakes testing and second chances: From data to models* (Order No. 31242034). Available from ProQuest Dissertations & Theses Global; Publicly Available Content Database. (3090764246). Retrieved from https://www.proquest.com/dissertations-theses/high-stakes-testing-second-chances-data-models/docview/3090764246/se-2

Sibley, C., Herdener, N., Coyne, J., Drollinger, S., & Strong, K. (2023, June 2). Exploring practice effects in the Navy's aviation selection test [Poster presentation]. 22nd International Symposium on Aviation Psychology. Rochester, New York, United States.

Stanek, K. C., & Ones, D. S. (2023). Meta-analytic relations between personality and cognitive ability. *Proceedings of the National Academy of Sciences*, *120*(23), e2212794120.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245-251. https://doi.org/10.1037/0033-2909.87.2.245

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric monographs*.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior research methods*, *37*(3), 498-505.

Whitehead, P. S., Brewer, G. A., & Blais, C. (2019). Are cognitive control processes reliable?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(5), 765.

Xu, Z., Adam, K. C. S., Fang, X., & Vogel, E. K. (2018). The reliability and stability of visual working memory capacity. *Behavior Research Methods*, *50*, 576-588.